1979

Holistic Essay Scoring: An Application of the Model for the Evaluation of Writing Ability and the Measurement of Growth in Writing Ability Over Time

Judith A. Powills
Loyola University Chicago

HOLISTIC ESSAY SCORING: AN APPLICATION OF THE MODEL FOR THE EVALUATION OF WRITING ABILITY AND THE MEASUREMENT OF GROWTH IN WRITING ABILITY OVER TIME

by

Judith A. Powills

A Thesis Submitted to the Faculty of the Graduate School of Loyola University of Chicago in Partial Fulfillment of the Requirements for the Degree of Master of Arts

April

1979

ACKNOWLEDGMENTS

The outstanding cooperation from my committee members, Dr. Jack Kavanagh and Dr. Ronald Morgan is gratefully acknowledged. The encouragement and stimulation they provided throughout the program deserves special recognition. The cooperation of the Indianapolis Public School staff including Ms. Roberta Bowers, Dr. Paul Brown, and Ms. Helen Cartwright was essential to the study and is deeply appreciated. Special thanks are also extended to the many Educational Testing Service staff who provided valuable advice and support including Dr. Jayjia Hsia and particularly to Ms. Gertrude Conlan who initially introduced me to the holistic essay scoring method. Appreciation is further extended to my family and friends.

Finally, loving thanks are extended to the very special person in my life. Without his unceasing support, this study would not have been possible.

VITA

The author, Judith A. Powills, is the daughter of Lee T. Goodman and Gladys L. Goodman and the wife of Michael A. Powills, III. She was born on August 12, 1951 in Milwaukee, Wisconsin.

Her high school diploma was received from Nicolet High School in Milwaukee, Wisconsin in 1969. In May, 1973 she received the degree of Bachelor of Arts with majors in psychology and social work from the University of Wisconsin in Madison.

Ms. Powills was employed by the Educational Testing Service Midwestern Regional Office in Evanston, Illinois in June of 1973 where she began her career as a Research Secretary. She was promoted to a Research Assistant in 1974 and became a Professional Associate for the company in 1978.

The author's publications include:

Holistic Essay Scoring: An Application of the Model for the Evaluation of Writing Ability and the Measurement of Growth in Writing Ability Over Time, (with G. Conlan and R. Bowers). Paper presented at the AERA Annual Meeting, April, 1979.

Evaluation of the Kellogg Company's Nutrition Education Program for Grades Seven Through Twelve, (with N. Samors). Educational Testing Service, Midwestern Regional Office, Evanston, Illinois, 1979.

WTTW/ESAA Television Pilot, "As We See It-2:" Formative Evaluation, (with T. Strand). Educational Testing Service, Midwestern Regional Office, Evanston, Illinois, 1978.

The Development of a Culturally Fair Model for the Early Identification and Selection of Gifted Children, (with T. Storlie, D. Bellis, and P. Prapuolenis). Educational Testing Service, Midwestern Regional Office, Evanston, Illinois, 1978.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

Table Page

CHAPTER I

THE PROBLEM

Educational evaluation has received increased emphasis in recent years. Since the late 1950's numerous essays have been written on program evaluation and conferences, workshops, speeches, and books have surged to address the issue. Program evaluation is in response to the concerns of education consumers and to the providers of funds and grants. Programs, spanning a wide range of subject areas, focus upon the knowledges, skills and abilities achieved within those subject areas. The 'back to basics' movement has encouraged evaluations of programs of mathematics, reading and writing at all levels of education.

Providing teachers with information about the teaching of writing, in particular, has received extensive interest. Such informational resources are applied within the classroom instructional framework; determination of the value and effectiveness of these informational resources as applied to a writing program requires product data available through the use of assessment instruments and methodologies. Such instruments and methodologies provide a key for decision making and the application of evaluation findings. One such assessment technique applied to the evaluation of writing ability is the holistic method of scoring student composition. If this method of analysis is to be of maximum effectiveness in terms of program evaluation, it is necessary to observe the method and its application in a model case study. To

that end, this study investigated the appropriateness and utility of using the holistic essay scoring technique in order to evaluate a writing program and to measure growth in writing ability over time.

Statement of the Problem

Examining a piece of student writing for its whole effect provides one means of analysis. Holistic scoring involves judging a composition for the total impression it creates, rather than judging individual components of writing including for example, spelling, punctuation and organization. A single score is given for the complete paper rather than a series of scores on particular aspects of writing skill. Scorers do not ignore separate facets of writing; they however, consider each factor as it relates to the whole. Performance on one facet does not control the judgment of performance on the whole. Through the method, teachers view writing as an integrative process; the various components of writing are considered simultaneously and teachers are encouraged to view teaching writing as more than the teaching of parts.

The method has been incorporated for national testing programs implemented by the Educational Testing Service which require examinees to respond to an essay examination topic. Following highly standardized and rigorous procedures, high reader reliability has been obtained. This is of particular relevance as proponents of the indirect and objective measurement of writing (e.g. multiple-choice instruments) argue their side in view of the most serious deficiency of essay examinations; the unreliability of the scoring. The utility of holistic essay scoring for national testing programs has been well documented (Smith, 1976).

The utility of the method applied in different situations such as within a classroom, a school or a district, however, merits study.

This study sought to determine the utility and outcomes of a holistic essay scoring in-service session for teachers of writing at the seventh- and eighth-grade levels. The compositions scored during the in-service session included pretest and posttest essays. The problems examined for this study were: 1) whether or not the in-service scoring session complied with the theoretical model of holistic essay scoring (i.e. judging a paper in relation to other papers within the population rather than against a preconceived ideal); 2) whether or not one could have confidence in the obtained scores (i.e. maintaining standards set by the readers and achieving consistency in scoring the essay papers as evidenced by reasonably high interrater reliabilities); 3) whether or not writing ability improved over progressive grade levels (i.e. finding significant differences in test performance of seventh- and eighth-graders); 4) whether or not writing improved over a short period of time (i.e. finding significant differences between pretest scores, test written in December and posttest scores, test written in the following April); 5) whether or not the writing program implemented by the study schools was effective (i.e. students of program participating teachers performing better at posttest time than students of non-program teachers); and finally, 6) whether or not the obtained essay scores were valid measurements of writing ability.

Need for the Study

The application of the holistic method of scoring essays is a

response to the demand for the direct measurement of writing ability. Writing abilities have most often been assessed by indirect measures such as standardized tests in a multiple-choice format. The obtained high reliability and validity coefficients, as well as high correlations with direct measures of writing lend credence to the use of such indirect measures. Nevertheless, teachers of writing have proclaimed that indirect measures provide analogous assessments of the students' ability to write, at best. If learning to write compositions is the ability to be measured, then actual composition it is said, should be assessed rather than some analogous behavior.

Direct measurement of writing surely, will continue to be sought particularly within the classroom; it will not give way totally to multiple-choice tests of writing skill. In fact, given a highly reliable and valid means for scoring composition, this direct measurement of writing may have stronger implications within the schools and within the district, for large-scale testing programs, and for the analysis of writing abilities for placement, admission, and employment purposes.

The need for this study is inherent in providing a quick, efficient, reliable and valid means to evaluate writing. Use of the holistic method by teachers of writing from varied backgrounds and different schools, their reactions to the technique, the results and implications of the scores obtained, can provide a basis for educators interested in the direct measurement of writing. Since the holistic method of scoring composition may provide a means of helping teachers to improve their teaching of writing, and of helping administrators to assess the writing skills of students in their schools, it would be useful to study the

use of the method by teachers of writing in order to evaluate composition, to measure growth in writing over time, and to evaluate a program designed to assist students in refining their knowledges and skills in the writing process. The methodologies and findings of this study may prove useful to other teachers and administrators with programs, goals, and objectives designed to improve student writing abilities.

The Study's Relationship to the Indianapolis Public Schools' Writer's Clinic

The Writer's Clinic was an in-service program for junior high school teachers implemented by the Indianapolis Public Schools and sponsored by a grant from the Lilly Endowment. Designed to aid teachers in enhancing the writing skills of their seventh- and eighth-grade students, the program was offered to all junior high teachers on a voluntary basis. Forty-eight Language Arts teachers of grades seven and eight participated and committed themselves to complete a series of activities during the course of thirteen in-service days of the Writer's Clinic in a one academic year period. The teacher contracts included: writing once a week with the students; guiding students in self-editing and self-proofreading; maintaining student folders containing writing; directing students in completing check sheets for the first composition of the year and the third composition of each six weeks period; submitting one flawless piece of prose from each student at the end of the project; and communicating with parents regarding the activities of the clinic and individual student progression (Bowers, 1978). Ultimately, in the final two days of inservice for the 1977-1978 Writer's Clinic,

pretest and posttest essays would be evaluated holistically not only to provide teachers with experience in the method, but also to evaluate the program's effectiveness in terms of student growth in writing.

The Writer's Clinic for teachers began with an intensive series of five work sessions during the summer. Session leaders included Indianapolis Public School professional and administrative staff (Ms. Roberta Bowers, Writer's Clinic Clinician and Ms. Helen Cartwright, Supervisor of Language Arts) as well as invited resource leaders from other institutions and organizations (Mr. Harvey Jacobs, Editor, The Indianapolis Star; Dr. H. Thompson Filmer, Professor of Education, University of Florida, Gainesville). Areas of concern for the initial in-service activities included: the starting point and setting of the project; composition through literature; our language today; uses of expository paragraphs in speech and debate and in reading units; the development of unit themes from unit themes to thesis statements; thought and language models in organizing experiences; and evaluation by writing, editing content, and proofreading. The primary focus of these early in-service days was on using a variety of models in the teaching of composition.

Two in-service work sessions including the topics of a writing experience from planning to final copy, editing and evaluating, sharing successes, and the basics of writing, were held in October and November. The second clinic was conducted by Dr. Hans Guth, coauthor of the series of English textbooks, American English Today. Dr. Guth demonstrated how composition could be taught through a progressive sequence from a word to a sentence to a paragraph and finally extending to a complete theme.

The third clinic was led by Ms. Roberta Bowers. In this session, participants were led through a lesson in descriptive writing. The outcome of this lession was participant-written essays that were compiled into booklet form.

In January, the Clinic included two work sessions directed by Dr. Michael Flanigan, Director of First Year Studies, Indiana University, Department of English. Topics included: a nine-step writing process; the skills of writers and the skills of teachers; and students as their own editors. Using specified models, particularly a process approach to writing which concentrated on the skills which writers possess and what writers actually do, participants focused on structure and audience and analyzed their own writing behaviors.

One work session comprised the fifth Writer's Clinic. Staff from the Indianapolis Public Schools including an evaluation consultant, Mr. Carl Hines, led the session. Participants discussed writing as a form of self-expression and covered the topics of reading, reflecting, and writing; preparing an entry for a Writer's Fair; and the evaluation of writing. The last segment of this work session in March, involved a sharing of successful experiences among the participating teachers.

The sixth Writer's Clinic meeting was conducted by Dr. Harrison J. Means, Associate Professor of English, Curriculum and Foundation Division, Cleveland State University. Dr. Means suggested strategies for teaching writing to slower students and discussed miscue analysis of writing and what it tells the writer, the diagnostic/prescriptive model format for teaching composition, criterion-referenced evaluation strategies, and sentence combining strategies.

The seventh and final meeting of the Writer's Clinic was held in May for two days. Teachers participating in the clinic activities, having incorporated their clinic experiences into their own classroom instruction, along with program administrators, were in need of an efficient method for evaluating the program in terms of its effectiveness as demonstrated by growth in the students' writing ability. Pretest and posttest essay papers had already been submitted for evaluation purposes. This concern was the focus of the last clinic where the Educational Testing Service conducted a two-day holistic essay scoring session. Clinic participants were trained in the method of holistic scoring and proceeded to read and score approximately 6,000 student compositions from two groups: experimental (students whose teachers had participated in the Writer's Clinic program) and comparison (students whose teachers had not participated in the program, but who were identified to represent characteristics of the experimental group). The methodology of this scoring session and the scores assigned to the students' compositions by the readers are the prime focus of the present investigation. Evaluation results of the present study and other assessment data were to provide a measure of the Writer's Clinic program effectiveness for the Indianapolis Public Schools.

Limitations of the Study

Because this study was subject to several limitations, caution in interpreting the results of the investigation is necessitated.

The study sample was not randomly selected. Experimental students who submitted pre- and posttest essays were seventh- and eighth-

grade students of teachers who had participated in the Writer's Clinic. The Writer's Clinic, in its first year of operation, was volunteer-participation-based. Therefore, the lack of sample randomness was a necessary limitation; it would neither have been feasible or desirable to have teachers assigned at random to participate in the clinic. It is also important to point out that the teachers had a written contract to comply with the goals, objectives and activities outlined for the program and described earlier in this chapter. This commitment, as well as participating in thirteen actual days of inservice, emphasized that the volunteering teachers were highly motivated. They can not be necessarily assumed to be a typical group of teachers.

Comparison group students were to be selected so as to be as similar as possible to student participants in the Writer's Clinic with respect to variables such as age, grade, sex, class, and school; while a one-to-one match was not necessary, the overall group profiles should have corresponded. While the comparison group was formulated to match characteristics of 700 randomly selected experimental students, the representativeness of the comparison group was less than ideal. This was due partially to the fact that comparison group students were selected from intact classes. Again, without working with intact classrooms, selection of any comparison students would not have been feasible. Additionally, comparison students were selected after the pretest essay had been administered to experimental students. While the comparison group students did respond to the same posttest essay assignment at the same time as experimental students, pretest essays were not received from comparison students. Keeping the design limitations in mind, analyses

including comparisons between treatment and non-treatment group post-test scores were conducted.

Overview of Thesis

Chapter I has included a statement of the problem, the need for the study, its relationship to the Indianapolis Public Schools' Writer's Clinic program and indication of the study's limitations.

Included in Chapter II is a review of the literature relevant to the study and the evaluation of writing ability in general. Chapter III concentrates on the hypotheses tested, the study design, methodologies and analyses. The results of the statistical analyses performed on the variables measured and discussion of them are presented in Chapter IV. Finally, Chapter V summarizes the study results and presents conclusions and implications for further research in the area of using the holistic method for the evaluation of student writing.

CHAPTER II

REVIEW OF THE LITERATURE

One present day concern of educators is that of the writing ability of the nation's students at all levels of education. The decline in the command of the English language and its communication through writing is confirmed through the 'back to basics' movement, discouraging press releases, and increased writing programming at the elementary, secondary, and post-secondary levels. A growing body of evidence supporting the writing decline has been summarized in The Concern for Writing (Educational Testing Service, 1978, pp. 1-2).

> Between 1963 and 1977, verbal scores on the SAT (Scholastic Aptitude Test) have dropped 49 points. While the verbal section of the standardized test is not a direct measure of writing ability, it does reflect one fundamental of written communication - facility with the language.
>
> Between 1970 and 1974, seventeen-year-olds' command of writing mechanics declined as reported by the National Assessment of Educational Progress. The students lacked the ability to organize ideas in written form, showed a tendency to write random sentences with simple vocabulary, and when faced with revision, corrected mechanical errors while neglecting to revise faulty organization.
>
> The 1977 study conducted by E. C. Ladd, Jr. and S. Martin Lipset for The Chronicle of Higher Education indicated that college faculty members agreed that students were not performing adequately in written and oral communication.
>
> The reading level of a pamphlet for college freshman prepared by the Association of American Publishers was altered from twelfth-grade to ninth-grade to promote student understanding.
>
> A commission funded by the National Endowment for the Humanities and formed by the Council for Basic Education is investigating the

writing crisis. A report is expected in 1979.

Remedial English courses are now being required for approximately one half of the freshman at the University of California at Berkeley because of great deficiency in writing ability.

The City University of New York recently mandated that students, before becoming juniors, must pass tests on a twelfth-grade reading level, and other tests on a ninth-grade math level in addition to writing an acceptable essay of 200 to 300 words.

The stress on performance in writing has broader implications for writing is an aspect of communication skills. At the strongest extreme, James Sledd notes the price to be paid for general literacy. He views standard English as being not only the creation and instrument of power and means of social control, but also as the vehicle of culture and means to self-development (Feinberg, 1978). Phrased more mildly, the Commission on English (1965) states that what is most important in speaking is also more important in writing - thought and its expression.

The many and varying hypotheses for the reasons for the decline in writing ability from television to advertising, to less emphasis in grammar and composition, to overloaded teachers, are not the focus of the present study. Still, a well-justified and direct response to the decline is the recognition that new programs are being designed to strengthen the writing abilities of the country's students. One of the most difficult aspects however, of any program is that of its evaluation and the evaluation of the abilities for which the program is designed to heighten. One development in response to college interest in, and demand for assessing writing ability is the reintroduction of the essay section to the College Board's Achievement Test after a six-year absence (College Board News, 1978). Beyond the scope of national

standardized testing, elementary and secondary schools in addition to colleges and universities, are emphasizing teacher inservice and program evaluation in the area of writing skill.

Some uncertainty regarding the evaluation of writing exists. As Cooper and Odell (1977) note, the early research and field work has been uncoordinated and agreement on standard procedures has not been reached. Nevertheless, a great deal of activity aimed at writing ability promises to have an effect for most importantly, the problem has been recognized.

Evidences of declining abilities and public concerns have resulted in a new education specialization - evaluation. Through the particular developments and innovations within the realm of evaluation, the importance of student test performance has remained constant. Pupil test performance plays a vital role in any approach to evaluation.

> Consequently, the more that educators find themselves obliged to evaluate their programs, the more concern will be given to the testing operation, since test results constitute such a key component of almost all educational evaluation. To employ the wrong tests for such an important task would be foolhardy (Popham, 1978, p. 4).

Direct Versus Indirect Evaluation of Writing

The need for evaluation of writing abilities and programs designed to improve those abilities has been recognized. Assessing the abilities remains crucial to the evaluation and hence, the need for appropriate assessment techniques cannot be underestimated. Several schools of thought pertaining to the assessment of writing ability exist. These schools can be basically divided into two groups: proponents of the direct measurement of writing and proponents of the indirect measurement of writing ability. Discussion however, of these types of

measures is interrelated for the advantage of one method is a disadvantage of the other. Clearly, such advantages and disadvantages of both methods adds to the state of confusion and the lack of agreement on standard procedures for the assessment of writing skill.

Indirect measures of writing are typically standardized or teacher-made examinations of the multiple-choice format. Measurement relies upon the correlation between performance on a test and performance on an actual test of writing. Direct measures on the other hand, are those which require the student to respond to a topic or question in essay or composition form: the student is asked to perform the actual task to be measured.

Proponents of the use of indirect measures argue that the weaknesses of essay tests are sufficiently significant. The most serious deficiency of the direct measure is the unreliability of essay scoring (Popham, 1978). Reliance on the subjective judgment in essay scoring reduces scoring reliability and test reliability is therefore, limited not only by the length of the test but also the scoring reliability (Conlan, 1978). Many studies have shown that different raters score the same written responses differently and that this scoring variability reduces obtained reliability coefficients (Klein & Hart, 1968). Such unreliable scoring leads to an unreliable, and therefore, invalid test.

Cooper and Odell (1977) state that direct measures are preferred over standardized tests which although widely used, generally only measure editorial skills including choosing the best sentence, recognizing correct usage, and incorporation of proper punctuation and capitalization. The editors further state that where writing samples are not

feasible to obtain, a simple measure of verbal ability would be preferred over a standardized test of editorial skills. Additionally, such standardized tests fail to assess the ability to compose for different purposes and audiences; a key element in the art of written communication.

Indirect measures are viewed by some as analogous assessments of students' ability to write, at best. It is suggested that if students learning to write composition is the ability to be measured, then composition through direct measurement should be assessed rather than some analogous behavior (Cohen, 1973). The direct essay measure increases face validity, for by requiring the actual task, that which can be measured is extended. Breland's (1977) analysis however, indicated that a short multiple-choice test of writing ability predicted actual writing performance as well as, or better than, a brief essay test during the freshman year of college. The relationship between direct and indirect measures was also exemplified by high correlations between the Test of Standard Written English scores and scores on actual writing samples indicating that the multiple-choice examination is not a measure of a behavior merely analygous to writing (Breland, 1977).

The skills measured by an essay examination are unlimited and total whereas in multiple-choice questions, the measurement is limited and fractioned (Conlan, 1978). Students responding to an essay topic or question must actually compose, organize, supply evidence, spell, and punctuate. In this respect, all aspects of writing can be measured at the same time. Not all aspects of writing can be measured via the multiple-choice format and writing skill is separated into parts that

are measured independently. For the most part, components of writing that can be measured in a single sentence receive the greatest emphasis.

Mechanics in terms of time required, sampling, the method of time for scoring, and costs favor the multiple-choice examination of writing skill over the essay examination (Conlan, 1978). Depending upon the item types used, multiple-choice test items can require as little time as thirty seconds per item where typically, a minimum of twenty minutes is required for response to an essay question. Because of such time factors, the sampling achieved by essay item types is limited. A maximum of three writing samples may be obtained in an hour. Therefore, the examinee who may not understand one question or who misinterprets the question misses on a major portion of the test. Indirect measures, on the other hand, can incorporate as many as 100 items per hour of examinee response time. Likewise, the candidate who incorrectly responds to one question is not in serious jeopardy.

The scoring methods for direct versus indirect measures of writing also favor the later. Essay questions must be individually scored by trained readers whereas indirect questions may be machine scored. Machine scoring costs and time are significantly less than the increased costs and increased time required for scoring essay questions.

Despite the strong criticisms given the direct writing measures, there are strong arguments in their favor in addition to some cited above. The logic stated by Coffman (1971) that a scholar performs by speaking and writing and that the essay examination constitutes a sample of scholarly performance and is hence a direct measure of educational achievement, is emphasized. In addition to an essay examination being

a direct measure of writing ability, essay items are often recommended as they provide practice in writing skills and give students an opportunity to improve writing (Popham, 1978). Also, despite Breland's (1977) findings, most professors involved in his study felt that actual student writing samples in addition to Test of Standard Written English scores were necessary in order to accurately place students in freshman English courses. English faculty generally approve of the essay item type and often react to the multiple-choice items with hostility and distrust (Conlan, 1978). Many English teachers believe that writing is much more than reducing writing to the level of subject-verb agreement as evidenced by the limits inherent in multiple-choice testing. Additionally, English faculty view multiple-choice tests as no more than exercises in error hunting and that the way to determine whether an individual can write is to have that person actually write. Essay questions can have an influence on the school curriculum and they are thought to encourage the requirement of actual writing in the schools. The method of teaching should be viewed as teaching students to write compositions rather than encouraging students to perform exercises in detecting errors.

Critics of the direct measure argue basically in terms of reliability and validity, but when the essay is used for program evaluation purposes rather than for individual student diagnosis and evaluation, and when exact, standardized procedures are followed, reliability and validity are conceptualized differently (Popham, 1978).

Analytic Versus Impressionistic Composition Scoring Procedures

Actual student composition will continue to be a means for the

assessment of writing ability and it is important to determine how to grade, mark or score the composition submitted, whether it be for a classroom assignment or as a test to be analyzed in terms of program evaluation. Methods for scoring lie on a continuum ranging from total analytic to total impressionistic scoring, both often serving different purposes. In analytic scoring, the crucial elements of an ideal answer are typically identified and scored more or less separately (Ebel, 1972). Attention is paid also to the relationships between the ideal elements forming the organization and integration of the answer. It is typically analytic scoring that results in a paper being decorated in red ink: spelling errors are circled; punctuation marks are inserted; and abbreviated comments such as 'awk' fill the margins when thoughts are expressed in an awkward manner. Impressionistic scoring generally involves looking at a composition in terms of its total effect. While emphasis too, is placed on certain elements of writing (punctuation, capitalization, spelling, diction), it is the interrelationship between these parts in terms of the whole, final product that is of particular concern. The holistic method of scoring student composition fits neatly into the impressionistic mode.

Neither method, analytic nor impressionistic scoring, is an end to the assessment of writing ability. There are advantages and disadvantages to both, and within each are variations. The usefulness of any method however, is dependent upon the type of composition to be analyzed and the goals and objectives to be measured (Fowles, 1977).

While in analytic scoring precise scoring criteria are provided, such is not the case with impressionistic scoring. At times, these

scoring criteria may seem restrictive and impressionistic scoring (the simultaneous consideration of various writing components) may be preferred. Analytic correcting of papers is important for certain purposes but has additional disadvantages. If teachers continue to score composition in isolation by decorating a paper with red marks in correcting mechanical and sentence errors, student morale and a positive attitude toward writing may be inhibited; a feeling that one has nothing acceptable to communicate permeates (Cooper & Odell, 1977). Stevens (1973) indicates results of a study where students who received only positive commentary on their papers developed a more positive attitude toward writing than those receiving negative commentary. On the other hand, the Commission on English (1965) states that whenever a student composition is read, it is the teacher's responsibility to mark errors in punctuation, grammar, spelling and diction even when only a most cursory reading is feasible. Ideally, the Commission further states that detailed comment should be made in addition to the marking of formal errors. Comments should be both constructive and specific and that which has been successfully accomplished should be stressed. One study (Jerabek & Diederich, 1975) suggests that the tone in which the teacher responds to student writing is a key issue. Further suggested was that the teacher, when written commentary is the means of response, should indicate two to three things which the student had done well, along with one suggestion for improving on the next assignment.

Impressionistic scoring, unlike analytic scoring, does not concern itself with marking errors. Rather, a grade, number, letter or single word is noted to either score the paper as a whole or is supplied

for each of the principal topics of ordinary correction of writing skills: punctuation; spelling; grammar; diction; organization; reasoning; and content. This impressionistic scoring is often a solution for the teacher who lacks the time for discursive comment or analytic marking. For most teachers, the ideal kind of reading including comment and marking is out of the question. Impressionistic scoring then, can save hours of correcting, and at the same time benefit the student especially when used in conjunction with analytic scoring.

Perhaps one of the best arguments in favor of impressionistic or holistic scoring is that of Feinberg (1978) who refers to W. Ross Winterowd's belief that the discipline underlying composition is rhetoric and rhetoric in itself is holistic. Rhetoric embraces reading and writing and is holistic. Therefore, the evaluation of our reading and writing should be holistic. The ability to write reflects the ability to communicate, and holistic scoring brings us closer to the essence of communication. This is substantiated by Cooper and Odell (1977) who state that at present, holistic evaluation gets us closer to what is essential in communication: writing is a communication with a whole message with a particular tone. Still, there are many variations of the impressionistic method of scoring composition to be considered.

Variations of Impressionistic Composition Scoring Procedures

The evaluation of writing in an impressionistic manner lies on a continuum from atomistic to holistic. While in holistic scoring, a piece of writing is considered to be a sample of a whole entity, in atomistic scoring particular features associated with skill in writing are assessed

(Cooper & Odell, 1977). In atomistic scoring, student composition is scored in terms of organization, diction, or spelling, or combinations of these and other rhetorical aspects are nominally judged in total. The identified features are known to be elements of discourse as a whole, however, these same features are isolated from the whole and are scored separately. Feature analysis is one method of atomistic scoring which focuses on a particular aspect of writing and is often used for descriptive writing tasks. Atomistic scoring is typically easier to use, is cheaper, and is sometimes more reliable than holistic scoring (Cooper & Odell, 1977). Nevertheless, holistic scoring which relates features of writing performance to discourse as a whole, is potentially more valid than atomistic scoring.

Holistic scoring assumes that a valid test of discourse requires an examination of a sample of discourse as a whole rather than merely as a collection of parts (Cooper & Odell, 1977). Still, holistic scoring often lends itself to a particular kind of writing sample and it cannot be said that excellence in one mode of writing can predict achievement in other modes of writing as well.

Primary-trait scoring, developed for the scoring of essays from the Second Assessment of the National Assessment of Educational Progress, is one method that takes into consideration that a writer of a clear technical report may not be able to produce an equally clear and persuasive letter to a congressman (Cooper & Odell, 1977). The rater's attention is focused on only those aspects of a piece of writing that are relevant to the kind of discourse required. For this method, scoring guides are developed for a particular writing task. In primary-trait

scoring, performance of a particular mode is stimulated by restricting the situation. This increases the chance that a certain writing exercise may in fact, fall outside the realm of a student's experience. Typically for true holistic scoring results reflect not only skill in manipulating language forms, but also reflect a student's experience in the required mode of response.

Several scoring scales fitting the impressionistic scoring mode predominate the literature including analytic, dicotomous, and essay scales. Analytic scales list prominant characteristics in a particular mode. Each feature is described in detail with high, middle, and low examples identified and described. The analytic scale is often used for placement or summative evaluation purposes.

The dicotomous scale, where a rater answers a series of statements with a simple 'yes' or 'no' as to whether the writing has the feature identified in the statement, is often sufficient for making gross distinctions between the quality of a series of compositions. When a series of compositions are arranged according to quality on a scale from inadequate to exemplary, an essay scale is incorporated. Raters attempt to place a new piece of discourse along this scale by matching it with the scale piece most similar to it.

Another variation of the essay scale, and perhaps the most simple procedure for scoring student writing, is general impressionistic scoring. This type of scoring is not quite as confining as those mentioned above. In this method of scoring, scores given to particular features of writing are not summed and a detailed discussion of such features is not assumed. Raters simply determine where a particular piece of writing fits within the range of writing produced for the assignment.

Without clear discussion however, of the features of writing, it is difficult to achieve inter- and intrarater reliability. Global quality scoring (Ebel, 1972) is one example of general impressionistic scoring. Here, the rater reads the composition for a general impression of its adequacy. The compositions are sorted into batches corresponding to different levels along a quality scale. Such a system is often beneficial to teachers of writing at the classroom level, for the rater is encouraged to reconsider his or her ratings in light of the experience with reading all of the students' responses. This can lessen the probability that if two papers seem to be of equal quality, one would receive a higher score.

The benefits of the use of the direct measure of writing as compared to the indirect measure, in combination with the method of scoring direct measures of writing require consideration of efficiency, reliability, and validity of the measure itself and the means by which the measure is scored. Objectivity in scoring is the key and just as there are different schools of thought concerning the type of measurement sought, so too are there different schools of thought concerning scoring systems. The present study was designed to demonstrate standardized procedures of holistic essay scoring over analytic scoring for the purpose of enhancing the validity and reliability of writing program evaluation.

The Validity of Direct and Indirect Measures of Writing

The validity of an assessment instrument in simple terms refers to the extent to which the instrument measures that which it purports to measure. Arguments in favor of the direct measurement of writing skills

(e.g. essay examination) stress stronger validity than that obtained via an indirect measure of the same skills. Validity of an instrument can take into account predictive, concurrent, content, and/or construct validity.

Numbers that demonstrate a degree of relationship typically obtained between sets of scores from two different measurements represent validity coefficients. The validities of standardized tests are often expressed in terms of these coefficients. The two different measurements relate to predictive validity. Here, one measure is designed to predict performance on the identified ability at some time in the future, or is designed to correlate with an identified criterion. A measure of writing skills, whether indirect (standardized multiple-choice test) or direct (essay examination) is often correlated with an accepted contemporary criterion of performance on the variable (writing skill) that the instrument is designed to measure (Ebel, 1972). One accepted criterion related to concurrent validity and a test of writing ability is an English course grade (Cooper & Odell, 1977). Breland (1977) however, cautions against the use of such a criterion because English course grades often represent more than writing ability and it is reasonable to assume they will correlate less highly with a true measure of writing ability.

Content validity is of particular value to program evaluation. Such validity is not arrived at through obtaining a validity coefficient; rather, an instrument is compared to program objectives in a judgmental fashion (California State Department of Education, 1977). Content validity of a writing skills instrument will ideally reflect an appropriate measure for a particular writing program; the instrument will measure what the writers have been practicing in the program (Cooper & Odell,

1977). A measure with true content validity concerning writing skills will reflect a comprehensive idea of written products and will also reflect the range of written products subsumed in the writing program (Cooper, 1975). The question to which content validity provides an answer is: Does a given measurement scheme permit writers to demonstrate what they have achieved in the course of instruction within the program (Cooper & Odell, 1977)?

Another type of validity is construct validity. If a measure of writing skills has construct validity, it will reflect an adequate reflection of the composing process; a construct of interest is measured (writing ability or writing performance). The key question in terms of construct validity is: Does a given measurement scheme permit a description of writing or the use of writing strategies (Cooper & Odell, 1977)? Teachers and researchers interested in the measurement of writing are primarily concerned with construct and content validity. While an assessment instrument may have valuable predictive validity, its content and construct validity may leave much to be desired. Teachers of writing often view tests of sentences, tests on rhetorical information, teacher-made tests of usage, spelling, punctuation, and capitalization, and standardized tests on usage rules, vocabulary, and syntax as exemplifying this lack of content and construct validity in terms of writing (Cooper, 1975). In measuring growth in writing in particular, the teacher and researcher are attempting to determine what happened rather than what will happen in the future.

Proponents of the direct measurement of writing which allows for an actual sample of student writing, continue to argue that the direct measure is clearly a more valid measurement of writing skill than is

multiple-choice testing, despite the high correlations between direct and indirect measures. Another critical issue closely related to validity concerning the assessment of writing is the reliability of the assessment instrument. As discussed earlier, those who favor the indirect measurement of writing argue primarily in terms of scoring reliability difficulties thought to be inherent in the essay examination. However, Cooper and Odell among numerous others claim that "where there is commitment and time to do the work required to achieve reliability of judgment," the essay-type examination used to evaluate writing skills, provides the most valid and direct means (Cooper & Odell, 1977).

The Reliability of Direct and Indirect Measures of Writing

An instrument's reliability demonstrates the extent to which it is consistent in measuring what it is intended to measure. There are many factors which can cause unreliability in a test including factors inherent in the test itself, factors which affect the individual taking the test, and factors connected with marking, grading or scoring the test (Stalnaker & Stalnaker, 1934). Coefficients of reader reliability are not to be confused with coefficients of examinee reliability nor test reliability (Ebel, 1972).

Factors in the test itself which cause unreliability are particularly common to objective, multiple-choice tests. Such factors include ambiguous items and defects in the test's mechanical makeup. Also included in this area of test reliability is how similarly the examinees perform on different, but closely equivalent tasks (Ebel, 1972). An individual taking a test takes with him/her a physiological state and a degree of motivation to the test situation. Each of these, along with

external physical conditions can cause reduced test reliability. A reliable instrument is fair to the person taking the test and allows him/her to do his/her best and to perform on one occasion similarly to performance on another occasion. Ebel (1972) refers to these factors as examinee reliability: how consistently the examinees perform on the same set of tasks. The third group of factors, those associated with the scoring of the test, are of direct interest to the present study. Objective, indirect measures of writing ability are often said to be scored with higher reliability than essay-type, direct measures. There is extensive evidence that consistency, or reliability, is difficult to achieve in scoring essay examinations (Coffman, 1971). Reader reliability refers to how closely two or more readers agree in rating the same set of papers and how consistently or inconsistently one reader rates a series of papers (Ebel, 1972). Scoring inconsistencies are identified by two distinct types of reliability: intra- and interrater reliability.

Intrarater reliability is determined by whether the same reader assigns the same score to the same compositions on different occasions (Klein & Hart, 1968). Studies have shown (Coffman, 1971) that a single rater does tend to assign a different grade to the same paper when it is reread at a later point in time. Interrater reliability refers to the tendency of different raters rating the same paper differently: different raters will tend to assign different grades to the same paper. A paper judged high by one rater may be judged low by another (Godshalk, Swineford & Coffman, 1966). Incorporated in this source of unreliability is the tendency of some raters to consistently rate higher than others, where still others will distribute themselves among higher and lower ratings

(Coffman, 1971). These differences among readers exemplify differing opinions on what characterizes good writing. A correlation among multiple sets of ratings provides a measure of the reliability with which the papers were read (Ebel, 1972).

Another source of unreliability lies within students; there are differences in student writing quality from one topic to another. This source relates to inter- and intrareliability, for differences within and among readers tend to increase as an essay question permits greater freedom of response (Coffman, 1971). Cooper and Odell (1977) note that in order to test a student's ability to write in varying modes, multiple pieces of writing on multiple occasions are necessary. This is because syntactic patterns vary from mode to mode (dramatic writing, personal writing, business writing) and it is typically the best writers who display most variation. Reliability however, can be increased substantially as more essays written by the same student are examined (Klein & Hart, 1968).

One of the most important standards of any type of assessment is quality. Scoring reliability when implementing essay examinations is therefore, significantly affected by the reliability with which responses are read and scored: the agreement between the scores of different raters and the consistency shown in scoring by individual raters. The value of using the direct measurement scheme of writing ability cannot be underestimated. This value though, is directly related to procedures followed to achieve maximum efficiency, reliability and validity. Holistic scoring, a quick and efficient system, incorporates these standardized procedures to achieve maximum reliability and validity and is the focus of this study.

Maximizing Essay Scoring Reliability and Validity

Procedures Prior to Scoring

Reliable reading and consequent scoring is maximized by following guidelines beginning at the development stage of the essay question itself. To begin with, essay questions are formulated so as to require a definite and restricted type of response (Stalnaker & Stalnaker, 1934). Ten general guidelines for writing essay questions proposed by Conlan (1976) provide a starting point for maximizing reliability and validity in the direct measurement of writing skill:

1. The question should be clear: Students should not have to puzzel over the instructions. The topic is intended to test the ability to write the answer and not the ability to guess what the test maker intends. Besides, students have only a limited amount of time, time that should be spent writing and not analyzing unnecessarily.

2. The question should be as brief as clarity allows: Restatement may sometimes be necessary to avoid misunderstanding. But, then, perhaps one should consider whether the restatement should be used without the original because the restatement does not need additional clarification.

3. The instructions should be definite: Students should know what is required. For example: Discuss, citing specific examples from one novel; or Pay attention to the correct form of the business letter; or Be sure to use complete sentences.

4. Avoid questions requiring only a yes or no answer: For example: Do you agree? Where does the student go from there?

5. Average students should be able to write average answers to the question, and yet bright students should be able to show their brightness: A good topic permits the ranking of all students according to ability.

6. The vocabulary used and the concepts expressed in the topic should not be too difficult for the ordinary student to understand immediately: A difficult topic distinguishes only between the very bright and the rest of the population. Besides, difficult reading changes the test to a reading test.

7. The question should not call for cliches as answers: A topic worn out by overuse produces worn-out responses. On the other hand, some good questions merely twist cliches. For example: What's right with television?; In what ways are teenagers more conservative than the over-thirties?

8. The question itself should provide an organizing principle for the essay: For example: Briefly describe...and then analyze; Discuss your answer to this question, giving the reasons for your answer and citing specific examples to support those reasons.

9. The question should not elicit responses which affect either the writer's or the reader's judgment: Politics, racial issues, and other inflamatory topics are to be avoided. If a candidate writes on the wrong political figure (from the reader's point of view), the score is either too high because the reader is making up for his or her own bias, or too low because the reader has succumbed to that bias. Also to be avoided are topics that are dishearteningly dull. For example, the fifth essay on the greatness of the basketball coach is not scored on the same standard as the first. Readers are human; they do become bored.

10. The question writer should write out the answer expected and determine whether the question really calls for that answer. The question writer should also try to answer the question in the allotted time, just to see whether it is humanly possible to do so. The question should be revised in the light of any discoveries made.

Many factors can result in reader biases which can distort the grading process. For example, knowledge that a particular student has written a given composition may bias a rater's appraisal of that student's response (Popham, 1978). Certain procedures followed prior to an essay scoring should be implemented in order to eliminate the halo effect insofar as it is possible. The halo effect includes a tendency to give a paper a score that is higher or lower than deserved due to certain impressions of the student performing the task. To reduce this effect, essay responses should be evaluated anonymously. Anonymity of response is assured in such a way in the holistic scoring of the College Board's Advanced Placement Test in English (Smith, 1976). By covering the

candidate's name, school and community, a reader can "grade the essay answers without being prejudiced, positively or negatively, by knowledge or irrelevant information about the candidate" (Smith, 1976, p. 4). This further ensures that each answer is judged solely on its own merit. In the situation where pre- and posttest essays are being read at the same time, it is also wise to conceal the date or other indication of which administration the paper came from; pretest or posttest (Odell, 1976). Concealing such identifying information concerning the candidate and administration period will reduce the possibility that the halo effect or other biases will influence the assigned grade or score (Ebel, 1972).

Another caution prior to the essay scoring session is to place all essays in a random order so that all papers from the same grade level, school or course are not arranged sequentially. Procedures for the grading of the Advanced Placement Test in English follow this guideline as well. "All essay booklets are placed in random order before the reading begins so that booklets from a particular school will not be preserved in a single set, possibly distorting the grading" (Smith, 1976, p. 4). A random shuffling of papers prior to their grading is also supported by Hales and Tokar (1975) who explain that a block of superior responses depresses the grades or scores assigned to subsequent responses, and a block of poor papers enhances the grades or scores assigned subsequent papers. Random ordering of papers then, minimizes the influence of these blocks on the grading of following responses (Klein & Hart, 1968).

Multiple evaluations on each paper should be mandatory particularly within a holistic scoring system to maximize reader reliability.

Studies have indicated that at least two readings by two different readers should be given to each paper (Dressel, Schmid & Kincaid, 1952). One check on the objectivity and reliability of the grading process is independent grading. Ebel (1972) states that at least two independent readings without consultation between the two readers is necessary. The correlations between the obtained pairs of grades indicate the reliability of grading the questions. Because there is typically then, at least two readings per paper in holistic scoring, it is necessary to devise a system prior to the reading to code or conceal first readers' scores or grades from the second readers. Objectivity in scoring and hence reliability in scoring, is heightened in this manner. In grading the Advanced Placement Examination in English, Smith (1976) indicates that by completely masking all scores given by other readers, the halo effect is further eliminated.

Before the reading begins, it is also suggested that a system be devised so that the shift of papers from one reader to another reader preserves the mixed random distribution of papers mentioned earlier. In addition, readers receiving papers already scored at least once should receive papers read by several readers rather than by only one other reader (Smith, 1976).

Training the Readers

Before any scoring can begin by a group of readers, the readers must be carefully trained (Cooper & Odell, 1977). In holistic essay scoring, papers are judged in relation to each other rather than against a preconceived ideal. Raters are generally requested to look at the papers for what they are rather than for what they should be. Reliability

in scoring cannot be achieved when raters use an absolute standard of quality (Cooper & Odell, 1977). Readers are asked to take into consideration the writing task, the conditions under which the responses were written, the age and ability of the subjects, and the full range of the quality of papers. Obtained holistic scores in this sense, generally approximate the normal curve distribution where some papers receive the highest scores and some the lowest, while more will fall into the middle range (Conlan, 1976).

Readers are trained for holistic scoring to decide as a group what factors they will look for in determining an overall judgment of the papers (Diederich, 1966). The training stage is critical for reliable holistic essay scoring. In a study conducted by Diederich, French, and Carlton (1961), fifty three raters were asked to rate 300 two-hour compositions written by college freshman by sorting the papers into nine piles. Participants were not trained nor were they given standards or criteria for judging the compositions. Without the intensive training, the outcome was that ninety four percent of the papers received seven, eight, or nine of the nine possible grades, and the median correlation between readers was .31.

Length of training varies depending upon many factors. Raters for the Diederich et. al. study (1961) included ten English teachers and forty three other raters (social scientists, natural scientists, writers and editors, lawyers and business executives). Careful training in combination with raters from similar backgrounds can improve reliability to an acceptable level (Cooper & Odell, 1977). Cooper and Odell further state that a group of homogeneous raters trained in holistic scoring can

achieve close to perfect agreement in choosing the better of a pair of essays; scoring reliabilities in the high 80s and low 90s can be obtained on the summed scores from multiple pieces of a student's writing. Higher reliabilities due to homogeneity of raters is substantiated by Follman and Anderson (1967). Stalnaker and Stalnaker in a 1934 study demonstrated that rater reliability could be improved from a range of .30 to .75 before training to a range of .73 to .98 after training (Cooper & Odell, 1977). Increased reliability of scoring direct measurements of writing is related then, to the homogeneity of raters and their experience as trained raters.

Procedures to guarantee that grading standards are applied tightly are likewise followed in the grading of the Advanced Placement Examination in English. Through the rigorous training, readers can apply the use of agreed-upon grading standards and can apply these scoring criteria fairly to all papers. The accurate and uniform assessment of papers is achieved by spending from three to seven hours of the five to six day reading period in training the readers (Smith, 1976). The objective of the training is two-fold: two essential components (each reader's judgment in assessing the answers and the careful standards developed within and by the group of raters) are stressed during the training period.

During the training, readers set their standards and criteria by reading and scoring a series of sample papers identified to represent the range of writing ability. The raters are therefore, trained carefully and become calibrated to reach consensus by reading and discussing a large number of representative papers. The official reading does not begin until there is close agreement in scoring among the readers

(Stalnaker & Stalnaker, 1934); a group consensus in scoring is achieved. The group is made aware of discrepancies in scoring. Coffman (1971) points out that when readers are made aware of their discrepancies, they will tend to move their own ratings in the direction of the group consensus. Through the training process, the ratings of the group then become more reliable and comparable standards can be maintained during the reading.

The Group Reading Activity

Close monitoring of grading or scoring standards receives considerable effort throughout the reading, thus maximizing reliability of scores. The maintenance of standards follows closely the procedures used in grading the Advanced Placement English Examination as explained by Smith (1976, p. 4):

> The Table Leaders, working with groups of 3 to 10 Readers, independently grade booklets which have previously been read or, alternately, ask a Reader to reread booklets which he or she has previously read, perhaps the day before. In either instance, if too great a disparity exists between the two sets of scores, the Table Leader and Reader discuss the papers to resolve differences. Maintaining the grading standards helps to guarantee that no matter when a candidate's answer is read or by whom, it will, with high probability, receive the same grade. Each of the reading groups augments the program-wide procedures by practices which assure that candidates will receive an accurate and fair estimate of their demonstrated achievement on the Advanced Placement Examination.

Speed of reading and scoring as well as control of the fatigue factor also contribute to the reliability of scoring. The College Board experience with holistic scoring sessions indicates that essay questions will be graded more reliably provided the graders are encouraged to work rapidly; readers are encouraged to read quickly for the total impression and are discouraged from rereading papers (Diederich, 1966). Fatigue

from reading is almost inevitable. Myers and McConville (1966) noted that reliabilities drop towards the end of the reading task when there is a loss of vigilance due to the anticipation of completing the task. The study showed an equivalent drop in reliability regardless of how long the reading lasted. An effort to bolster reader morale and effort is especially critical at this point in the reading process. Frequent rest breaks are mandatory in conjunction with the maintenance of standards throughout the reading. Systematic error as a result of reader fatigue and anticipated task completion is somewhat counterbalanced by the implementation of multiple readings where all papers are read once before they are distributed for the second of subsequent evaluations.

By observing proper precautions as those discussed above, the reliability of reading essay tests can be achieved. The significant criteria of a test item, whether it is an objective or essay item, include the item reliability, validity and the fidelity with which it measures what it is intended to measure (Stalnaker & Stalnaker, 1934). Ebel (1972, p. 143) stresses that the value of a score is dependent upon its objectivity and reliability:

> To the degree that other qualified observers would assign different scores, the measurement lacks objectivity and hence, utility. If the same teacher were to assign totally different scores to the same essay test on different occasions, or if different teachers were to disagree in the same way, our confidence in the scores would be shaken and their usefulness diminished.

The Application of Holistic Essay Scores

Evaluation and research concerning student writing abilities has been concentrated in various areas: measuring students' growth in writing over time; determining the effectiveness of a writing program; measuring

group differences in writing performance in comparison-group research; describing the writing performance of individuals or groups in developmental studies; and studying hypothesized correlates of writing skill (Cooper & Odell, 1977). Different areas are often studied simultaneously; for example, obtaining measures of student growth in writing can lead to a scheme for writing program evaluation and assessing the program's effectiveness (Wagner, 1975). Holistic essay scores provide a measure for assessing growth and evaluating writing programs. The usefulness of scores is dependent upon their reliability as noted earlier. A key step then, in research analyses, must be consideration of the utility of the obtained measures before proceeding to analyses concerning writing growth or program evaluation. The Diederich study of 1974 indicated that reliability coefficients of .80 are adequate for program evaluation, .90 for individual growth measurement (Cooper & Odell, 1977).

Examining growth in writing and subsequent program evaluation must also take into account the course of the writing process itself. Writing is a complex skill that is neither taught nor learned in a short period of time (Breland, 1977). Beaven suggests that there are many factors that affect student growth in writing for improvement in writing does not occur in isolation; writing relates to speaking, listening, reading, and other areas of communication (Cooper & Odell, 1977). The slowness in the growth of writing is substantiated by others in the literature (Wagner, 1975). The Commission on English (1965) has reaffirmed the gradual acquisition of skill in writing and states that the growth is inseparable from the processes of physical, social, emotional, and intellectual development. Beaven cites the research confusion of the growth issue (Cooper & Odell, 1977). Many studies hypothesizing the efficacy of

various instructional methods often fail to show significant improvement in writing. Often, these studies have allowed only six, ten, or even fifteen weeks. Because improvement may occur over a much longer period of time, nonsignificant results are dubious. One reinforcement of the assumption that growth in writing occurs slowly is Cazden's study in 1972 which showed that children may take eight to nineteen or twenty months to achieve mastery once they have begun to use plural forms of nouns (Cooper & Odell, 1977).

Measuring Growth in Writing Over Time

The pretest-posttest research design incorporating holistic essay scores is often implemented in studying growth in writing ability over time. Gain scores from pre-administration to post-administration are often used for evaluation purposes. Caution in interpreting these gain scores however, is stressed in the literature. Gain scores are systematically related to any random error of measurement and these scores obtained by subtracting pretest from posttest scores often lead to questionable conclusions (Werts & Linn, 1970). The unreliability of these growth measures is due to the accumulation of measurement errors; each test score includes its own error of measurement and when it is subtracted from another measurement the errors accumulate rather than cancel out (Ebel, 1972). Apparent gain is a good measure of true gain only when the tests used are perfectly reliable (Lord, 1956).

Ceiling and floor effects are particularly related to gain scores (Diederich, 1966). The pretest might have been too easy for high scoring students and there is little room for improvement on the posttest primarily because their initial scores were so near to the maximum possible

scores (ceiling effect). Likewise, students with the lowest pretest scores appear to gain most at posttest time (floor effect) because they have a greater likelihood of showing larger gains than those who earned higher initial scores. Statistical methods preferred over raw gain scores include analyses of variance or the use of residual gain scores because of the inherent problems of raw gain scores (Cronbach & Furby, 1970).

Another difficulty with measuring growth in student writing from pretest to posttest involves the use of different topics or assignments between administrations. It cannot be guaranteed that the essay topics were of the same difficulty for the students responding at pretest and at posttest time and therefore, conclusive statements about the writing gains cannot be made. One alternative to control for this factor is to have both pre- and posttest essays scored at the same time by the same group of readers who are unaware of which topic was used for the pretest and posttest administrations (Breland, 1977). Ideally, the same topic should be used for both administrations to control for difficulty level provided enough time has lapsed between administrations to control for the practice effect.

In program evaluation there is concern for how the observed growth compares with expected growth without program treatment. Therefore, a reference group should be used to compare the growth behaviors of program participants to that of similar, but non-program participants. Reference groups can be of three types: control; comparison; or norm groups (California State Department of Education, 1977). A control group design involves randomly assigning subjects to the program (treatment) or non-program (no treatment) groups. Comparison groups include existing classes

of non-program participants that are comparable to program participants. Random assignment is not a requirement for comparison group usage and while participants and non-participants need not match on a one-to-one basis, the overall group profiles should be similar. Comparison group subjects should be given the same instruments on the same schedule as program participants. Norm reference groups are sometimes used when control or comparison groups are not feasible. They often however, represent a broader population and may not be comparable to the program group. Program evaluation, and the design components that are part of the evaluation is in summary, closely tied to growth measurement because the growth resulting from the program directly taps the effectiveness of the program.

Holistic Essay Scoring: Other Applications

In evaluating student writing, improvement of writing, and the effectiveness of a writing program, the direct measurement of writing via an essay test provides an appropriate means of assessment. The validity of the essay for writing assessment has received great attention over standardized, indirect measures of writing skill and rigorous procedures for scoring student composition holistically can significantly increase the direct measure's reliability. Still, holistic scoring has other applications in addition to its use for measuring writing growth and evaluating writing programs.

The assignment of essays for in-class or out-of-class assignments, in addition to classroom testing and testing for research purposes is encouraged when the teacher desires to encourage and reward the development of student skill in written expression (Ebel, 1972). Ebel further notes that essay tasks encourage the cultivation of written expression and

provide practice in it. Providing more writing experiences for the student provides a partial answer to correcting deficiencies in writing ability. Studies often suggest that if students are given an increased opportunity to exercise writing skills, then improvement in expression through writing may be a natural consequence. Wagner (1975) expands the argument by suggesting that writing ability will necessarily remain undeveloped if practice is not provided. Without practice, advancement and maturation is unlikely.

> Given more opportunity to practice various forms of writing, to experiment with new forms and ideas and to revise writing, the greater the opportunity will be for the student to gain insight into himself, language, and skill with written communication (Wagner, 1975, p. 77).

This insight can result in a mature and competent command of English expression. The benefits of practice in writing are further confirmed by findings of a study conducted at Purdue University by Locke and Wykoff (Arnold, 1964) which showed that students doing twice as much writing as other students showed fewer failures and greatest improvements.

By following the guidelines established by the Commission on English (1965) calling for detailed comment in addition to marking errors on each student paper, the teacher is often prohibited from assigning essay topics; little practice in writing for the student is provided. The long and dull hours of grading papers leads teachers to shy away from assigning tasks that require students to organize and express themselves in written prose (Wagner, 1975). Microscopic grading, viewed by many as not even helpful to the student, has caused a disappearance of the weekly writing assignment.

Holistic scoring, when used to supplement comprehensive grading, provides classroom teachers an opportunity to give more writing assignments,

perform classroom evaluation, and at the same time, provide feedback to students. This further provides students the opportunity to practice and subsequently, improve their writing abilities.

Recapitulation

This chapter has included a review of the literature concerning the need for writing programs and the evaluation of these programs. Writing skill can be measured by different types of assignments ranging on a continuum from direct to indirect measurement. The direct measurement of writing, while often more time consuming than indirect, maintains credibility and worth due to its validity; students are required to actually compose by organizing and developing a response which provides a definite demonstration of writing skill. Reading and scoring reliability, which appear to be the major drawbacks of the direct measure, can be maximized by the holistic approach to scoring following a set of clearly documented procedures. At the same time, massive numbers of student compositions can be scored quickly and efficiently. When reliability is maximized, one can have confidence in the resulting scores which have meaning not only for the classroom teacher of writing, but also for the evaluation of a program and the measurement of growth in writing ability over time.

The value of essay scores has been shown to be irrefutable.

> Many believe the best way to measure such essentials as organizing ability, clarity of expression, and other more intricate and subtle factors is to have students write. Evaluating writing samples is not, however, a precise process since it depends heavily on human judgments. Because of a general lack of confidence in subjective evaluations in this objective age, and because of the length of time it takes to score writing samples, essays have not been used much in large-scale testing programs in recent years (ETS, 1978, p. 8).

However, where there is commitment to the valid and reliable measurement of writing performance, it can be achieved. The holistic method of evaluation provides one solution to the problems of scoring student writing and the validity and reliability of directly measuring writing skills.

> All the activity aimed at improving writing ability promises to have an effect. But probably not quickly. Writing is a demanding endeavor and one cannot learn to do it well overnight. But the problem has been recognized and many things are being done about it (ETS, 1978, p. 16).

CHAPTER III

METHODOLOGY

The purpose of this study was to investigate and analyze the scoring results of a holistic essay scoring session implemented to judge student compositions written for a program evaluation. Student compositions scored during the session included pretests written in December of 1977 and posttests written in April of 1978 by seventh- and eighth-grade students of the Indianapolis Public Schools. The opportunity to conduct the study occurred in connection with the Writer's Clinic developed by the Indianapolis Public Schools and funded by the Lilly Endowment.

This chapter includes a description of the questions to be answered, the study groups, data collected, the design and experimental procedures of the study, a discussion of the instruments used and procedures used to conduct the scoring session and to measure the variables under study, a statement of the hypotheses tested, and discussion of the statistical techniques used in analyzing the data.

Questions to be Answered

The questions to be investigated included:

1. Does the case study fit the theoretical model of holistic essay scoring?
2. Is consistency in scoring the essay papers achieved as evidenced by determination of interrater reliabilities?

3. Is there a significant difference in test performance of seventh- and eighth-graders as indicated by the obtained total test scores?

4. Is there significant growth in writing ability over time, from pretest to posttest?

5. Is the Writer's Clinic in-service program for teachers effective as indicated by a significant difference in posttest scores obtained by the comparison and experimental groups?

6. Are the pretest and posttest obtained scores valid measurements of writing ability?

Sample

The final sample for analyses included 5,788 compositions written by 4,071 seventh- and eighth-grade students in the Indianapolis Public Schools. Eighty-eight compositions were considered invalid and were deleted from the analyses. Invalid compositions included those that were off the topic or were incomprehensible or those for which student identifying information was missing.

Study subjects were from two groups; comparison (no treatment) and experimental (treatment). The comparison group, formed after experimental students were administered the pretest essay assignment, was administered a posttest only. Of the 4,071 study students, 648 formed the comparison, posttest only subgroup.

The experimental group was comprised of three distinct subgroups: students who had taken the pretest only; students who had taken the posttest only; and students who took both the pretest and posttest (matched pairs). The pretest only and posttest only experimental subgroups resulted due to mobility or absence on the day of pre- or posttest administration. Of the 3,623 experimental students, 1,717 were in the matched

pairs subgroup; 804 were in the pretest only subgroup; and 902 formed the posttest only subgroup.

Comparison group students came from eight schools, experimental students came from thirty-one schools across subgroups. Tables 1 and 2 show the numbers and percents of students in the comparison and experimental subgroups respectively, by school and by grade. Of the 648 comparison group students, 27.62 percent were from the seventh-grade and 72.38 percent were from the eighth-grade. The distribution of students by grade among the three experimental subgroups was 53.64 percent seventh-graders and 46.36 percent eighth-graders. The balance of grade by group (comparison and experimental) was less than ideal. Had the comparison group been selected to represent the characteristics of the experimental group, closer percentages of seventh- and eighth-graders across treatment groups would have been expected.

Table 1

Distribution of Comparison Group (Posttest Only) Students by School and Grade (N=648)

School	Seventh-Grade		Eighth-Grade	
	N	%	N	%
1	51	40.2	76	59.8
2	--	--	68	100.0
3	60	45.8	71	54.2
4	--	--	65	100.0
5	25	48.1	27	51.9
6	25	100.0	--	--
7	--	--	106	100.0
8	18	24.3	56	75.7
Total	179	27.6	469	72.4

Table 2

Distribution of Experimental Group Students by Subgroup, School, and Grade (N=3,423)

School	Pretest Only				Posttest Only				Pretest/Posttest Matched Pairs			
	7th		8th		7th		8th		7th		8th	
	N	%	N	%	N	%	N	%	N	%	N	%
1	3	37.5	5	62.5	17	58.6	12	41.4	24	43.6	31	56.4
2	18	52.9	16	47.1	12	50.0	12	50.0	10	31.3	22	68.8
3	13	39.4	20	60.6	20	54.1	17	45.9	66	56.4	51	43.6
4	15	100.0	--	--	11	100.0	--	--	56	100.0	--	--
5	21	55.3	17	44.7	15	44.1	19	55.9	44	56.4	34	43.6
6	31	64.6	17	35.4	24	85.7	4	14.3	16	94.1	1	5.9
7	6	46.2	7	53.8	9	45.0	11	55.0	17	35.4	31	64.6
8	--	--	5	100.0	--	--	15	100.0	--	--	27	100.0
9	4	33.3	8	66.7	12	44.4	15	55.6	8	25.8	23	74.2
10	--	--	27	100.0	--	--	--	--	--	--	--	--
11	7	100.0	--	--	15	100.0	--	--	52	100.0	--	--
12	22	71.0	9	29.0	15	45.5	18	54.5	33	45.2	40	54.8
13	11	100.0	--	--	29	100.0	--	--	42	100.0	--	--
14	1	16.7	5	83.3	5	38.5	8	61.5	22	45.8	26	54.2
15	3	37.5	5	62.5	3	75.0	1	25.0	15	62.5	9	37.5
16	17	100.0	--	--	21	100.0	--	--	74	100.0	--	--
17	55	100.0	--	--	--	--	--	--	--	--	--	--
18	27	100.0	--	--	15	100.0	--	--	12	100.0	--	--
19	5	100.0	--	--	3	100.0	--	--	21	100.0	--	--
20	19	52.8	17	47.2	15	53.6	13	46.4	42	45.2	51	54.8
21	2	12.5	14	87.5	8	53.3	7	46.7	18	30.5	41	69.5
22	11	100.0	--	--	34	100.0	--	--	34	100.0	--	--
23	23	40.4	34	59.6	28	58.3	20	41.7	50	42.4	68	57.6
24	--	--	27	100.0	--	--	44	100.0	--	--	30	100.0
25	--	--	8	100.0	33	78.6	9	21.4	--	--	37	100.0
26	--	--	2	100.0	4	30.8	9	69.2	17	50.0	17	50.0
27	49	52.1	45	47.9	3	60.0	2	40.0	18	42.9	24	57.1
28	13	44.8	16	55.2	13	27.1	35	72.9	70	56.0	55	44.0
29	--	--	10	100.0	--	--	8	100.0	--	--	36	100.0
30	32	42.7	43	57.3	132	55.0	108	45.0	97	40.6	142	59.4
31	29	74.4	10	25.6	9	47.4	10	52.6	36	57.1	27	42.9
Total	437	54.4	367	45.6	505	56.0	397	44.0	894	52.1	823	47.9

Determination of Differences Between Treatments

Writer's Clinic participants were interested in directly evaluating student writing ability. The program clinician developed an essay topic to be administered to all seventh- and eighth-grade students whose teachers volunteered for clinic participation. This essay assignment served as the pretest for the evaluation of writing abilities. The students were to respond to the same essay topic at the end of the school year. This second assignment constituted the posttest for the evaluation. A pretest-posttest design for the evaluation of writing skill was therefore, implemented.

In order to determine the effectiveness the teacher in-service program had in terms of student writing ability, the study sample was comprised of two primary groups. The group taking both the pretest and posttest was the experimental group. Treatment for this group consisted of the classroom experiences incorporated by the students' teachers who had volunteered to participate in the Writer's Clinic program. The 'no treatment' group, comprised of students whose teachers did not participate in the Writer's Clinic formed the comparison group. Comparison group students, unlike experimental students, did not write in response to the pretest essay assignment. Both groups, experimental and comparison, were compared using t-tests of differences between posttest means.

Dependent Measures

Data collected for the study included pretest and/or posttest scores assigned the student essays written for the Writer's Clinic evaluation. The scores were assigned during a holistic essay scoring session

conducted in May of 1978. This session represented the final Writer's Clinic activity for the school year. Each paper was read and scored by three different readers during the session. A total of forty four of the fifty Writer's Clinic participants assisted in the scoring. After scoring discrepancies were resolved, the three individual scores given each paper were summed to derive a total paper score. Student identifying information including school and grade level, were collected and matched with first, second, third and total reading scores.

In order to obtain some measure of concurrent validity, the essay scores were correlated with a criterion derived from class composition grades and teacher ratings of overall student writing ability. Each of the Writer's Clinic teachers who participated in the scoring session were asked to provide student information to reflect the criteria for students in one of their classes. A random assignment schedule was arranged to determine the classes for which criteria information would be provided. The teachers were given a form and were asked 1) to enter the names of all students in the identified class, 2) to enter the grades given the essay assignments for each month from December to May corresponding to each student, and 3) to rate each student on his/her overall writing ability in December and again in May using a scale of one (low) to four (high). Forms were received from only fifteen of the forty four teachers. Nine of the fifteen teachers taught students at the seventh-grade level and six taught students at the eighth-grade level. Class essay grade data were collected for 396 students across the classes and grade levels; 244 seventh-graders and 152 eighth-graders. These data were matched with the pretest and/or posttest scores obtained during the

essay scoring session. Table 3 shows the numerical distribution of students whose scores were matched with the criterion data (grades).

Three of the fifteen teachers who completed the form did not enter the overall ratings of writing ability for their students but did supply essay grades; two eighth-grade classes and one seventh-grade class lacked the overall rating data. Overall ratings were obtained for a total of 300 students.

Table 3

Numerical Distribution of Students With Matched Essay Test Scores and Composition Grades

Grade	Grades & Pretest	Grades & Posttest	Grades & Both Tests	Grades & No Tests	Total
Seventh	23	51	135	35	244
Eighth	14	32	80	26	152
Total	37	83	215	61	396

In order to determine the relationships between composition grades and essay test scores and teacher ratings and essay test scores, only those students having both pretest and posttest essay scores were included in the analysis (N=215). Of the 215 students, 135 were eighth-graders and eighty were seventh-graders. The teachers reported composition grades but did not rate fifty seven of these students in terms of overall writing ability in December and May (three classes); overall ratings were received for 158 of the 215 students.

The number of composition grades submitted varied by class and by student. No student had more than five composition grades reported

within any one month (December through May). Because the grades were separated into two periods relative to the pretest period and posttest period (December to February and March to May) students having fewer than four reported grades within either of the periods were deleted from analysis. Means for reported grades were completed for both periods (pre and post) and it was determined that students with fewer than four reported grades within a three month period could bias the period mean.

Description of the Holistic Essay Scoring Procedure

Activities Prior to the Scoring Session

Students were requested to respond to a direct measure of writing ability; they were to respond to a given topic by developing a composition theme. Concerned with measuring progress in writing effectively, it was decided that two controlled writing experiences would be conducted, one in December (pretest) and one at the end of the school year (posttest in April). The controlled writing assignment incorporated several procedures: there were no prewriting activities; the assignment was given succinctly and a certain amount of time was allowed for the writing (thirty minutes); students were instructed to use standard junior high school format; students were allowed to revise and rewrite during the allotted time; teachers were not able to assist except to clarify the directions; the assignments were not to be corrected or graded by the teachers; and teachers were given a limited time period in which to have students complete the assignment.

The difficulty level of the topic was controlled for by using the same topic for the two test administrations. Research has shown that

different essay topics do in fact, represent different difficulty levels (Godshalk, Swineford, & Coffman, 1966). With the variability of different topics controlled for, reading standards would remain consistent between pretest and posttest compositions. The practice effect was considered to be minimal because of the four month time lapse between the test administrations.

The essay topic selected for the pretest and posttest essay assignments was developed by a professional writing clinician of the Indianapolis Public Schools. Established guidelines for writing appropriate essay topics were followed. The essay topic and directions for its administration appear in Appendix A.

Student pretest and posttest compositions were submitted to the Writer's Clinic clinician for processing previous to the scoring session. In preparation for the reading, all papers were randomly mixed among the classrooms, teachers, grade levels, schools and test type (pretest or posttest).

Anonymity of student responses was assured by secluding all identifying information (name, grade, class, teacher, school, test date). Removeable labels covered this identifying information. By assuring writer anonymity, reader bias was controlled for; the reader would be unaware of who the writer was, what class, grade or school the writer was from, and whether or not the paper was written for the pretest or posttest administration.

Prior to the scoring session, workshop leaders read the majority of pretest and posttest student compositions in order to select sample papers that would be used for training in setting scoring standards. Sample papers were selected to represent the various levels (ranges) of

population writing ability from excellent to average to poor. The score scale that was to be used for the session ranged from one to four (low to high, respectively). Sample papers included those judged to be representative of each score. The number of samples was normally distributed across scores; there were more samples scored by the workshop leaders as 'two' and 'three' than samples scored as 'four' and 'one.' Other sample papers selected were those that might present some other interest to the readers. Sample papers are presented in Appendix B.

Unique score code cards were developed for the readers to assure independent judgment during the reading. These score codes were established to keep the judgments of each of the three readers of each paper independent. Each reader was to receive a unique score code so that when the second reader read a paper, he or she would be unable to interpret the score assigned by the first reader, and likewise for the third reading. The master score code appears in Appendix C.

Approximately twenty papers were inserted into each of many manila file folders; readers would score a complete folder before scoring papers in another folder.

Scoring Session Procedures

Introduction to holistic scoring. Forty four teachers of the Indianapolis Public Schools were in attendance at the scoring session. The first task for the workshop leaders was to introduce the method to the readers. Readers were told that holistic scoring is impressionistic scoring. The paper is read and scored for the total impression it creates rather than for particular aspects of writing skill such as punctuation, organization, diction, and/or spelling. The score scale of 'one' to

'four' was explained to the readers. In a scale which has only four points, there is obviously no middle score. One of the reasons why the score scale works as well as it does, is that readers are forced away from a middle or uncommitted score. Readers were told they would make two judgments for each paper: first, a decision was required as to whether the paper was in the upper half or lower half (above average or below average); then, readers were to decide if the paper was good enough to be rated as the highest score, or poor enough to be rated as the lowest score. Papers were to be judged in relation to the other compositions in the population rather than against a preconceived ideal. How the papers would be judged would be determined through standards set by the group. Readers were to take a positive approach to rating the compositions. They were asked to concentrate on what the student had accomplished rather than on what the student had failed to do or had done badly. Readers were also reminded that the students had only thirty minutes in which to respond to the topic, and that they were unable to use aides such as a dictionary.

Setting the standards. Following the introduction to holistic scoring, scoring standards were set through the examination of the selected sample papers. Sample papers were distributed to readers one-by-one. Participants were asked to read the composition quickly, and to score it on a scale of 'one' to 'four' for the total impression it created. The workshop leader then asked for an open vote on how many people had assigned a score of 'four,' of 'three,' of 'two,' or of 'one,' to the paper. Group votes were counted, tallied, and recorded on a blackboard visible to the entire group. A total of six sample papers

representing the range of writing abilities were distributed, scored, and openly voted upon in this manner. After each of the six samples had been scored, readers were asked to reexamine these papers and to put them in rank order as they related to each other. This was to reinforce the stipulation that each paper be read and judged against the other papers, rather than against an ideal. Participants were then free to discuss the characteristics of any particular paper; readers who had assigned scores of 'four,' 'three,' 'two,' or 'one,' discussed the reasons for assigning the respective scores.

Examination of sample papers one-by-one and discussion of them by the group continued until the readers had reached consensus on the standards for judging the papers. Discrepancies in scoring were discussed. A discrepancy occurred if all readers had not judged an individual paper alike. For example, if the majority of readers placed a paper in the above average category (three to four), and other readers placed the paper in the below average category (one to two), discussion was called for. Discussion of the paper assisted the low scorers to view the paper as the majority viewed the paper. Divergence from the standards on the part of any reader was corrected. Readers were to conform to the standards set by the group. Whether or not consensus was reached was determined by the examination of the count of scores as recorded on the blackboard. After a series of papers had been read, consensus was achieved; the majority of readers assigned the same score to each sample paper. Adjacent score discrepancies (e.g. 'four'/'three') were allowed, but discouraged. Non-adjacent scores were viewed as discrepancies and were not allowed (e.g. 'four'/'two'). It is important to note that the workshop leader did not impose standards on the group; rather, the group set its

own standards.

Actual scoring procedures. After standards for the reading were set with consensus in scoring achieved as evidenced by the scores given to the string of sample papers, each reader was given his or her own unique scoring code. Readers were instructed as to the use of this score code, and how to write the code on the paper. They were instructed to mark first their reader number and then, their score code assignment. Each reader received a folder of approximately twenty papers and was requested to read each paper in the folder and to score it according to the standards set during the training. Upon completion of a folder, the reader was given another folder. During the first round of reading, each folder was checked by the workshop leaders to assure the score code, and appropriate score code was used. The workshop leaders also read a random sample of papers from each folder to assure that scoring standards had been maintained, and that the readers were adhering to these standards in their scoring. This also provided a check for intrarater consistency. Readers had been instructed to bring all problem papers to the attention of the workshop leaders. All papers were read once before distribution for the second round of reading.

Before papers were distributed for the second reading, they were taken out of the original folders and randomly mixed. New folders were compiled which included papers read and scored by many different readers during the first reading. This mixing system was incorporated so that the second reader would receive papers already read and scored for the first time by several readers. Through this system, the situation of the same pair of readers continually reading each other's papers did not

occur. The same mixing procedure was implemented after all papers had been read twice, before distribution for the third reading.

At the end of the first day of the scoring session, all papers had been read at least once, and a large proportion had been read two times. A mini-training session took place at the beginning of the second day. To assure that the standards set during the first day of the scoring session were maintained on the second day, sample training papers from the first day were intermixed with new sample papers as a reliability check. Likewise, to assure that standards set were maintained throughout the reading, sample papers were distributed periodically, particularly after a rest break had been taken.

By the middle of the afternoon on the second day of scoring, all papers had been read and scored three times by three different readers. Papers were redistributed to the readers to convert the score codes into actual scores. Readers also totalled the three individual scores to obtain the total score.

Resolution of Score Discrepancies

Following the reading, papers were examined by the clinician of the Writer's Clinic to resolve discrepancies in scoring. Discrepancies were defined as non-adjacent assigned scores. Examples of acceptable scores across the three readings included: 4-4-4; 3-3-3; 2-2-2; 1-1-1; 4-4-3; 3-3-4; 3-3-2; 2-2-3; 2-2-1; 1-1-2. Examples of discrepancies included: 4-4-2; 4-4-1; 3-3-1; 4-2-1. The clinician resolved each discrepancy so that the three scores were in fact at least adjacent.

A total of 258 papers of the 5,788 papers read and scored had

discrepancies (4.46 percent). Table 4 shows the breakdown of these discrepancies by group. The resolution of discrepancies involved changing one, two, or three of the scores assigned. Table 5 shows the number of scores per paper necessitating change by group. Most papers had only one score changed (84.11 percent); fewer needed a change for two of the three scores (15.12 percent); and only two papers (.78 percent) required a change for all three scores.

Table 4

Distribution of Pretest and Posttest Discrepant Scores by Group (N=258)

Group	Total Number of Cases	Discrepant Scores	
		Number	Percent
Comparison	648	29	4.47
Experimental Matched Pairs/Pretest	1,717	68	3.96
Experimental Matched Pairs/Posttest	1,717	90	5.24
Experimental Pretest Only	804	36	4.48
Experimental Posttest Only	902	35	3.88
Total	5,788	258	4.46

Table 5

Number of Scores Necessitating Change to Resolve Discrepancies Across Three Readings by Group

Group	Number of Scores Changed (Total Possible=3)		
	One Score	Two Scores	Three Scores
Comparison	25	4	--
Experimental Matched Pairs/Pretest	59	9	--
Experimental Matched Pairs/Posttest	69	19	2
Experimental Pretest Only	33	3	--
Experimental Posttest Only	31	4	--
Total	217	39	2

Table 6 shows where these changes were made: on the first reading score; on the second reading score; on the third reading score; or on combinations across the three readings. Clearly, most changes (35.65 percent) were made to the first reading score. Table 7 shows how the resolution of discrepancies affected the total test score (sum of the three reading scores). Only three cases of the 258 discrepancies resulted in no change to the total score. Total scores were decreased for 123 (47.67 percent) of the total discrepancies and were increased for 132 (51.16 percent) of the total. The range of changes to the total score was from minus three score points to plus five score points.

Table 6

Location of Changed Scores Across Three Readings by Group

Group	Location of Changed Scores						
	1st	2nd	3rd	1st and 2nd	1st and 3rd	2nd and 3rd	1st, 2nd and 3rd
Comparison	10	8	7	2	1	1	--
Experimental Matched Pairs/ Pretest	23	19	17	6	3	--	--
Experimental Matched Pairs/ Posttest	31	15	23	5	7	7	2
Experimental Pretest Only	12	7	14	1	2	--	--
Experimental Posttest Only	16	7	8	2	1	1	--
Total	92	56	69	16	14	9	2

Table 7

Distribution of Changes in Total Score Points as a Result of Discrepancy Resolutions by Group

Group	Change in Total Score Points								
	-3	-2	-1	0	+1	+2	+3	+4	+5
Comparison	--	1	14	--	9	4	--	1	--
Experimental Matched Pairs/Pretest	--	13	25	--	21	8	--	1	--
Experimental Matched Pairs/Posttest	2	9	29	3	32	12	1	1	1
Experimental Pretest Only	--	2	16	--	15	3	--	--	--
Experimental Posttest Only	--	1	11	--	13	9	--	1	--
Total	2	26	95	3	90	36	1	4	1

Preparation of Data for Analyses

The labels covering the identifying student information were removed so that each paper could be coded according to the following variables: student identification number; school code; grade; pretest first reading score, second reading score, third reading score, and total pretest score; posttest first reading score, second reading score, third reading score, and total posttest score. These variables were coded, keypunched and verified. All variable data were then computer edited for valid values. After errors found by the computer edit were corrected, a five percent random sample of cases (N=200) was generated to determine coding errors. The analysis identified a .4 percent item coding error rate.

Hypotheses

Theoretically, in a large-scale holistic essay scoring session, the assigned scores should distribute themselves normally. This is due primarily to the fact that the papers in such a scoring session are read in relation to each other rather than against a preconceived ideal. Therefore, for each of the three readings given each paper, it is likely that more papers will be assigned scores of 'two' and 'three' than those of 'one' and 'four,' when a four-point scoring scale is employed. The number of papers scored as 'one' and 'four' should be approximately equal as should the number of papers assigned scores of 'two' and 'three.' When the three individual reading scores are summed to determine the total score with a range of three to twelve, it is expected that total scores in the middle range of six to nine should outnumber the number of scores in the high and low extremes, three to five and ten to twelve, respectively.

H_1: The distribution of scores for each of the three readings and summed totals across pretest, posttest, seventh- and eighth-grades, comparison and experimental groups, and all schools, should approximate a normal distribution.

The value of the obtained scores is dependent upon their objectivity and reliability. To the extent that other qualified readers would assign different scores to the same compositions, the measurement lacks objectivity and utility. It is necessary then, to establish the usefulness of the scores and have confidence that they are in fact, reliable. Research has shown that it is difficult to achieve consistency in grading essay examinations: different raters tend to assign different grades

to the same paper; a single reader tends to assign different grades to the same paper on different occasions; and differences in scoring tend to increase as the essay question permits greater freedom of response (Coffman, 1971). Through the observance of proper standardized procedures however, the reliable reading of essay tests can be achieved. Such procedures to ensure as high a reliability as possible include the anonymity of student responses, the random mixing of papers across all variables, the use of multiple ratings by different readers, and the maintenance of comparable standards during the reading through close monitoring.

H_{2a}: The interrater reliabilities reflect consistency in scoring.

H_{2b}: The number of papers with discrepant scores is less than the number of papers with non-discrepant scores.

H_{2c}: The number of papers receiving the same score by each of the three readers (perfect agreement) is greater than the number of papers not reflecting perfect agreement.

The acquisition of writing skill is gradual. It is inseparable from the processes of physical, social, emotional, moral, and intellectual growth (Commission on English, 1965). Given the ideal educational system, however, it would be expected that there would be progression and acquisition of additional skills as the student proceeds through his or her educational experience. In this sense, it would be likely that students at a higher educational grade level would perform better than students at a lower educational level.

H_3: Eighth-grade students will show higher total essay scores than seventh-grade students within each of the pretest and posttest administrations.

As stated earlier, growth in writing ability occurs slowly. The research studies designed to measure the effectiveness of varying methods of instruction often show a nonsignificant improvement in writing. The method's effectiveness is not necessarily disproven by such nonsignificant results however, for often researchers allow too short a period of time to measure change (Cooper & Odell, 1977). Given a longer period of research time, improvements may be significant. The time span between the pretest and posttest administrations for the present study was close to a five month period.

H_4: Obtained pretest and posttest total scores reflect growth in writing ability over time within and across grade levels.

Measuring group differences in writing performance using comparison group research is a widely used technique for program evaluation; determining the effectiveness of a writing program or a teacher of writing. Such research studies must be properly designed and the limitations of the present study's design must be recognized. Writing is a complex task that is not easily taught in a short period of time even with a dedicated program of instruction (Breland, 1977). Growth in writing ability is related to many factors and experiences both within and outside of an English classroom: speaking; listening; reading; and other information processing avenues of communication (Cooper & Odell, 1977). Improvement in writing does not occur in isolation. Nevertheless, the impetus of the present study lies in the implementation of an in-service program for teachers of writing. Some measure of the program's effectiveness, or lack of it, is indicated by a comparison of the experimental (treatment) and comparison (no treatment) groups.

H_5: The obtained total posttest scores of experimental students will be greater than those of comparison students within and across grade levels.

The degree of validity of a writing sample depends upon the kinds of tasks put to the writer and is closely tied to the reliability of the instrument. If a measure has construct validity, it reflects an adequate conception of the composing process. Construct validity is also based upon the agreement and consistency among those who read and judge the composition. If the test is valid, the readers will agree in their ranking of the essays, and their judgments would be stable. Content validity indicates that an appropriate measure has been employed; it permits the writer to demonstrate what he/she has achieved. While teachers and researchers are primarily interested in content and construct validity, concurrent validity has also been recognized in order to relate test scores to an accepted, contemporary criterion of performance on the variable that the test is intended to measure. For the present study, an effort was made to determine the concurrent validity of the pretest and posttest essay assignments by correlating student test scores with the essay grades teachers assigned to a sample of the student population during the course of their instruction from pretest to posttest time. Another criterion, teacher overall judgments of student writing skill at the beginning of the term and at the end of the term, was related to the pretest and posttest essay scores. Construct and content validity of the pre- and posttest essay assignments was established: the Writer's Clinic administrative personnel felt an essay assignment was a direct measure of student writing ability which has construct validity; content validity

was established by developing an assignment that was intended to evoke the best from the writer.

H_{6a}: There is a relationship between grades assigned to compositions at the beginning of the year and pretest holistic essay scores.

H_{6b}: There is a relationship between grades assigned to compositions at the end of the year and posttest holistic essay scores.

H_{6c}: There is a relationship between teacher ratings of students' overall writing ability at the beginning of the year and pretest holistic essay scores.

H_{6d}: There is a relationship between teacher ratings of students' overall writing ability at the end of the year and posttest holistic essay scores.

Null Hypotheses

The theoretical hypotheses stated above are restated in the form of null hypotheses as follows:

H_1: The distribution of scores for each of the three readings and summed totals across pretest, posttest, seventh- and eighth-grades, comparison and experimental groups, and all schools does not approximate a normal distribution.

H_{2a}: The interrater reliabilities do not reflect consistency in scoring.

H_{2b}: The number of papers with discrepant scores is greater than or equal to the number of papers with non-discrepant scores.

H_{2c}: The number of papers receiving the same score by each of the three readers (perfect agreement) is less than or equal to the number of papers not reflecting perfect agreement.

H_3: There are no differences in mean total essay scores of eighth-grade students and seventh-grade students within each of the pretest and posttest administrations.

H_4: There are no differences in mean total essay scores of the pretest and posttest administrations.

H_5: There are no differences in mean total essay scores of experimental and comparison students within and across grade levels.

H_{6a}: There is no evidence to support a relationship between grades assigned to compositions at the beginning of the year and pretest holistic essay scores.

H_{6b}: There is no evidence to support a relationship between grades assigned to compositions at the end of the year and posttest holistic essay scores.

H_{6c}: There is no evidence to support a relationship between teacher ratings of students' overall writing ability at the beginning of the year and pretest holistic essay scores.

H_{6d}: There is no evidence to support a relationship between teacher ratings of students' overall writing ability at the end of the year and posttest holistic essay scores.

Data Analyses

Data analyses for the study included t-tests between means and

analysis of variance. Statistics included means, standard deviations, the chi square statistic, Pearson correlations, and Cronbach's alpha coefficients of reliability. In all tests of significance, the .001 level of probability was used at the point at which the null hypotheses would be rejected. Values between the .05 and .01 levels were considered to be within the region of doubt. Additionally, the null hypotheses were accepted when values did not reach the .05 level of significance.

CHAPTER IV

RESULTS AND DISCUSSION

This chapter is divided into eight sections. Descriptive summary statistics for the sample of pretest and posttest student compositions are present first. Next, findings are presented in response to each of the questions of concern to the study: 1) Does the case study fit the theoretical model of holistic essay scoring? 2) Is consistency in scoring the essay papers achieved as evidenced via determination of inter-rater reliabilities? 3) Is there a significant difference in test performance between seventh- and eighth-graders as indicated by obtained posttest scores of the comparison group and obtained pretest and posttest scores by the experimental group? 4) Is there significant growth in writing ability over time (from pretest to posttest) within the experimental group? 5) Is the Writer's Clinic in-service program for teachers effective as indicated by a significant difference in posttest scores obtained by comparison and experimental groups? 6) Are the pretest and posttest obtained scores valid measurements of writing ability and what is the relationship between scores and classroom composition grades and teacher ratings of overall student writing ability? Finally, a summary of findings is reported.

Descriptive Summary Statistics

Each student composition, whether pretest or posttest, was read three times by three distinct readers on a scale of one to four with four representing the highest score. The total score for each paper was the sum of the three individual scores. The total possible raw score was twelve; the minimum possible score was three. The distribution of raw scores for the comparison group and the three experimental subgroups by reading, school and grade, is shown in Tables 8 through 12. Mean raw scores for the comparison group and the three experimental subgroups by reading across schools and grades are presented in Table 13.

Table 13 shows the mean posttest raw scores for the experimental subgroups on all readings to be consistently higher than those obtained by the comparison group. The table also shows the mean raw scores for the posttest on all readings to be consistently higher than those obtained on the pretest within the experimental group.

One-way analyses of variance were conducted to determine the existence of no differences among the means of the various schools by comparison and experimental subgroups. The analysis of variance results are presented in Tables 14 through 18. The null hypothesis that the schools do not differ with respect to total scores (pretest and/or posttest) was rejected at the .001 significance level for each of the comparison group and experimental subgroups. Because the F ratios were significant in each group, a posteriori tests could be conducted to determine where these differences between schools lie.

Table 8

Distribution of Raw Scores Earned on Posttest by Comparison Group Students by Reading, School and Grade (N=648)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1								
7th Grade	1.88	.77	1.76	.71	1.86	.66	5.51	1.75
8th Grade	1.88	.83	1.99	.68	2.05	.69	5.92	1.85
Total	1.88	.80	1.90	.70	1.98	.68	5.76	1.81
2								
8th Grade	2.43	.79	2.51	.82	2.50	.80	7.44	2.03
3								
7th Grade	2.25	.73	2.27	.61	2.22	.56	6.73	1.45
8th Grade	2.28	.80	2.24	.76	2.25	.71	6.77	1.91
Total	2.27	.76	2.25	.69	2.24	.64	6.76	1.71
4								
7th Grade	2.76	.52	2.84	.77	3.16	.62	8.76	1.33
8th Grade	2.89	1.01	2.78	.69	2.74	.66	8.41	2.12
Total	2.83	.81	2.81	.85	2.94	.67	8.58	1.76
5								
7th Grade	2.32	.96	2.64	.86	2.52	.71	7.48	1.92
6								
8th Grade	2.23	.81	2.24	.64	2.33	.73	6.79	1.71
7								
7th Grade	2.28	.89	2.22	.73	2.33	.78	6.83	2.07
8th Grade	2.16	.68	2.30	.71	2.27	.73	6.73	1.68
Total	2.19	.73	2.28	.71	2.28	.73	6.76	1.77
8								
8th Grade	2.29	.77	2.46	.71	2.38	.73	7.14	1.80

Table 9

Distribution of Raw Scores Earned on Pretest by Experimental Pretest Only Subgroup Students By Reading, School and Grade (N=804)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1								
7th Grade	1.33	.58	1.33	.58	1.33	.58	4.00	1.73
8th Grade	2.60	1.14	2.20	.45	2.00	1.00	6.80	2.17
Total	2.13	1.13	1.88	.64	1.75	.89	5.75	2.38
2								
7th Grade	2.22	.81	1.78	.65	2.11	.68	6.11	1.64
8th Grade	1.69	.60	1.69	.60	1.81	.40	5.19	1.11
Total	1.97	.76	1.74	.62	1.97	.58	5.68	1.47
3								
7th Grade	2.38	.77	2.62	.65	2.38	.87	7.38	2.02
8th Grade	2.05	1.05	2.10	.72	1.85	.75	6.00	2.20
Total	2.18	.95	2.30	.73	2.06	.83	6.55	2.21
4								
7th Grade	1.87	.74	2.20	.41	1.67	.62	5.73	1.39
5								
7th Grade	2.05	.80	2.10	.70	2.00	.71	6.14	1.90
8th Grade	1.71	.77	1.76	.56	2.06	.66	5.53	1.66
Total	1.89	.80	1.95	.66	2.03	.68	5.87	1.80
6								
7th Grade	2.10	.65	2.35	.75	2.48	.63	6.94	1.53
8th Grade	2.35	.86	2.82	.81	2.53	.80	7.71	2.20
Total	2.19	.73	2.52	.80	2.50	.68	7.21	1.81
7								
7th Grade	2.50	.84	2.50	.84	2.17	.75	7.17	2.23
8th Grade	2.29	.49	2.29	.76	2.14	.90	6.71	1.70
Total	2.38	.65	2.38	.77	2.15	.80	6.92	1.89
8								
8th Grade	2.20	.45	2.40	.55	2.20	.45	6.80	.84
9								
7th Grade	2.25	.96	2.25	.50	1.75	.50	6.25	1.71
8th Grade	2.38	.92	2.38	.74	2.38	.74	7.13	1.81
Total	2.33	.89	2.33	.65	2.17	.72	6.83	1.75

Table 9

(Continued)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
10								
8th Grade	3.26	.81	3.11	.75	3.15	.82	9.52	1.91
11								
7th Grade	2.00	.58	1.86	.90	2.14	.69	6.00	1.53
12								
7th Grade	2.36	.66	2.14	.71	2.36	.73	6.86	1.81
8th Grade	2.22	.97	2.22	.67	2.11	.93	6.56	2.40
Total	2.32	.75	2.16	.69	2.29	.78	6.77	1.96
13								
7th Grade	1.91	.70	2.00	.63	2.00	.63	5.91	1.81
14								
7th Grade	2.00	.00	2.00	.00	2.00	.00	6.00	.00
8th Grade	2.00	.71	2.40	.55	2.20	.84	6.60	1.82
Total	2.00	.63	2.33	.52	2.17	.75	6.50	1.64
15								
7th Grade	2.00	1.00	2.33	.58	2.00	.00	6.33	1.15
8th Grade	2.40	.89	2.80	1.30	2.80	.84	8.00	2.74
Total	2.25	.89	2.63	1.06	2.50	.76	7.38	2.33
16								
7th Grade	2.29	.92	2.35	.79	2.35	.79	7.00	2.21
17								
7th Grade	2.24	.84	2.40	.71	2.40	.68	7.04	1.83
18								
7th Grade	2.52	.89	2.37	.69	2.30	.67	7.19	1.80
19								
7th Grade	2.80	1.30	2.80	1.30	2.80	.84	8.40	3.29
20								
7th Grade	2.11	.81	1.95	.71	2.05	.62	6.11	1.73
8th Grade	2.35	1.00	2.35	.86	2.00	.71	6.71	2.20
Total	2.22	.90	2.14	.80	2.03	.65	6.39	1.96
21								
7th Grade	3.00	.00	2.50	.71	3.00	.00	8.50	.71

Table 9

(Continued)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
21								
7th Grade	3.00	.00	2.50	.71	3.00	.00	8.50	.71
8th Grade	1.71	.73	1.93	.62	1.93	.47	5.57	1.22
Total	1.88	.81	2.00	.63	2.06	.57	5.94	1.53
22								
7th Grade	2.18	.87	2.00	1.00	2.36	.92	6.55	2.62
23								
7th Grade	2.17	.94	1.87	.63	2.17	.72	6.22	1.70
8th Grade	2.21	.64	2.29	.68	2.21	.69	6.71	1.43
Total	2.19	.77	2.12	.68	2.19	.69	6.51	1.55
24								
8th Grade	2.81	.83	2.96	.81	2.74	.86	8.52	2.17
25								
8th Grade	2.88	.83	2.75	.71	2.50	.53	8.13	1.36
26								
8th Grade	4.00	.00	4.00	.00	4.00	.00	12.00	.00
27								
7th Grade	2.49	.92	2.51	.82	2.65	.81	7.65	2.17
8th Grade	3.09	.87	3.09	.63	2.91	.73	9.09	1.76
Total	2.78	.94	2.79	.79	2.78	.78	8.34	2.10
28								
7th Grade	1.92	.86	1.77	.60	2.00	.58	5.69	1.60
8th Grade	2.31	.79	2.31	.60	2.25	.68	6.88	1.67
Total	2.14	.83	2.07	.65	2.14	.64	6.34	1.71
29								
8th Grade	2.30	.82	2.70	.82	2.20	.79	7.20	2.20
30								
7th Grade	1.72	.68	1.78	.75	1.69	.59	5.19	1.65
8th Grade	2.44	.96	2.47	.77	2.37	.72	7.28	2.05
Total	2.13	.92	2.17	.83	2.08	.75	6.39	2.15
31								
7th Grade	2.03	1.02	1.97	.87	1.97	.82	5.97	2.43
8th Grade	2.40	.97	2.40	.97	2.50	.85	7.30	2.41
Total	2.13	1.00	2.08	.90	2.10	.85	6.31	2.46

Table 10

Distribution of Raw Scores Earned on Posttest by Experimental Posttest Only Subgroup Students By Reading, School and Grade (N=902)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1								
7th Grade	2.29	.85	2.24	.75	2.12	.60	6.65	1.84
8th Grade	2.08	.67	2.25	.75	2.25	.62	6.58	1.73
Total	2.21	.77	2.24	.74	2.17	.60	6.62	1.76
2								
7th Grade	2.75	.97	2.75	.97	2.83	.58	8.33	2.19
8th Grade	2.58	.51	2.92	.79	2.67	.78	8.17	1.64
Total	2.67	.76	2.83	.87	2.75	.68	8.25	1.89
3								
7th Grade	2.95	.89	2.75	.55	2.50	1.00	8.20	1.94
8th Grade	3.12	1.05	3.06	.90	3.41	.62	9.59	2.21
Total	3.03	.96	2.89	.74	2.92	.95	8.84	2.15
4								
7th Grade	2.27	.79	2.09	.83	2.45	.69	6.82	2.04
5								
7th Grade	2.27	1.10	2.20	.86	2.00	.65	6.47	2.26
8th Grade	1.79	.86	1.89	.66	1.95	.78	5.63	1.98
Total	2.00	.98	2.03	.76	1.97	.72	6.00	2.12
6								
7th Grade	2.08	.78	2.08	.65	2.29	.62	6.46	1.47
8th Grade	2.50	1.00	2.25	.96	2.00	1.41	6.75	2.99
Total	2.14	.80	2.11	.69	2.25	.75	6.50	1.69
7								
7th Grade	2.56	.73	2.33	.71	2.44	.53	7.33	1.22
8th Grade	2.27	.65	2.36	.50	2.18	.75	6.82	1.54
Total	2.40	.68	2.35	.59	2.30	.66	7.05	1.39
8								
8th Grade	2.73	.80	2.73	.88	2.33	.82	7.80	2.11
9								
7th Grade	2.08	.79	2.25	.87	2.33	.65	6.67	1.83
8th Grade	2.27	.80	2.13	.52	2.13	.64	6.53	1.46
Total	2.19	.79	2.19	.68	2.22	.64	6.59	1.60

Table 10

(Continued)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
10								
7th Grade	2.27	.80	2.27	.70	2.13	.64	6.67	1.72
11								
7th Grade	2.80	.77	2.73	1.03	2.80	.77	8.33	2.23
8th Grade	2.28	.89	2.22	.73	1.94	.73	6.44	1.82
Total	2.52	.87	2.45	.90	2.33	.85	7.30	2.20
12								
7th Grade	1.79	.73	1.93	.75	1.90	.72	5.62	1.84
13								
7th Grade	2.20	1.30	2.20	.84	2.00	.71	6.40	2.41
8th Grade	3.00	.93	2.63	.92	2.88	.83	8.50	2.56
Total	2.69	1.11	2.46	.88	2.54	.88	7.69	2.63
14								
7th Grade	2.00	1.00	2.67	1.15	2.67	.58	7.33	2.52
8th Grade	1.00	.00	3.00	.00	2.00	.00	6.00	.00
Total	1.75	.96	2.75	.96	2.50	.58	7.00	2.16
15								
7th Grade	2.29	.78	2.19	.81	2.38	.92	6.86	2.20
16								
7th Grade	1.80	.77	2.13	.74	1.80	.68	5.73	1.94
17								
7th Grade	3.00	1.00	3.33	.58	3.00	.00	9.33	1.53
18								
7th Grade	2.20	.41	1.87	.64	2.13	.52	6.20	1.21
8th Grade	1.54	.66	1.92	.64	2.00	.58	5.46	1.56
Total	1.89	.63	1.89	.63	2.07	.54	5.86	1.41
19								
7th Grade	2.88	.83	2.88	.83	3.13	.83	8.88	2.17
8th Grade	2.14	1.22	2.14	1.22	2.29	1.38	6.57	3.78
Total	2.53	1.06	2.53	1.06	2.73	1.16	7.80	3.14
20								
7th Grade	2.62	.92	2.71	.84	2.47	.79	7.79	2.20

Table 10

(Continued)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
21								
7th Grade	2.50	.92	2.36	.73	2.04	.69	6.89	1.97
8th Grade	2.70	.80	2.50	.83	2.40	.88	7.60	2.14
Total	2.58	.87	2.42	.77	2.19	.79	7.19	2.05
22								
8th Grade	3.27	.73	3.27	.73	3.27	.73	9.82	2.18
23								
7th Grade	2.70	.77	2.79	.78	2.58	.75	8.06	1.78
8th Grade	2.44	.73	2.89	.33	2.33	.50	7.67	.87
Total	2.64	.76	2.81	.71	2.52	.71	7.98	1.63
24								
7th Grade	3.00	.00	3.50	.58	3.25	.50	9.75	.96
8th Grade	3.78	.44	3.78	.44	3.78	.44	11.33	.71
25								
7th Grade	2.00	1.00	1.67	1.15	1.67	.58	5.33	2.31
8th Grade	3.00	.00	2.50	.71	2.50	.71	8.00	1.41
Total	2.40	.89	2.00	1.00	2.00	.71	6.40	2.30
26								
7th Grade	2.08	.64	2.46	.66	2.54	.66	7.08	1.32
8th Grade	2.46	.89	2.40	.74	2.31	.68	7.17	1.96
Total	2.35	.84	2.42	.71	2.38	.67	7.15	1.80
27								
8th Grade	2.63	.74	2.00	.00	2.13	.83	6.75	1.49
28								
7th Grade	2.90	.97	2.86	.91	2.78	.83	8.55	2.40
8th Grade	2.51	.85	2.55	.81	2.35	.80	7.41	2.00
Total	2.73	.94	2.72	.88	2.59	.84	8.03	2.30
29								
7th Grade	2.11	.78	2.22	.67	2.33	.87	6.67	2.06
8th Grade	2.30	.82	2.40	1.08	2.40	.84	7.10	2.23
Total	2.21	.79	2.32	.89	2.37	.83	6.89	2.11

Table 11

Distribution of Raw Scores Earned on Pretest by Experimental Pretest/Posttest Matched Pairs Subgroup Students by Reading, School and Grade (N=1,717)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1								
7th Grade	2.29	.62	2.13	.68	2.08	.58	6.50	1.53
8th Grade	2.23	.80	2.26	.68	2.35	.61	6.84	1.55
Total	2.25	.73	2.20	.68	2.24	.61	6.69	1.54
2								
7th Grade	2.10	.88	1.80	.79	1.90	.87	5.80	2.20
8th Grade	2.23	.87	2.59	.80	2.32	.78	7.14	2.03
Total	2.19	.86	2.34	.87	2.19	.82	6.72	2.14
3								
7th Grade	2.24	.77	2.39	.74	2.26	.66	6.89	1.69
8th Grade	2.52	1.00	2.45	.92	2.43	.78	7.41	2.49
Total	2.37	.89	2.42	.82	2.33	.72	7.12	2.09
4								
7th Grade	2.38	.95	2.19	.82	2.14	.75	6.71	2.15
5								
7th Grade	2.14	.88	2.14	.82	2.16	.81	6.43	2.10
8th Grade	1.82	.76	1.79	.59	1.88	.64	5.50	1.78
Total	2.00	.84	1.99	.75	2.04	.75	6.03	2.01
6								
7th Grade	2.50	.82	2.69	.70	3.00	.73	8.19	1.94
8th Grade	3.00	.00	3.00	.00	3.00	.00	9.00	.00
Total	2.53	.80	2.71	.69	3.00	.71	8.24	1.89
7								
7th Grade	2.65	.70	2.35	.60	2.35	.93	7.35	1.90
8th Grade	2.45	.72	2.52	.77	2.29	.74	7.26	1.77
Total	2.52	.71	2.46	.71	2.31	.80	7.29	1.80
8								
8th Grade	2.15	.82	2.22	.84	2.11	.75	6.48	2.14
9								
7th Grade	2.50	.93	2.50	.53	2.30	.76	7.50	2.00
8th Grade	2.52	.59	2.39	.72	2.22	.67	7.13	1.58
Total	2.52	.68	2.42	.67	2.29	.69	7.23	1.67

Table 11

(Continued)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
10								
7th Grade	1.83	.76	2.04	.74	2.00	.71	5.87	1.85
11								
7th Grade	2.58	.87	2.36	.60	2.55	.71	7.48	1.73
8th Grade	2.23	.77	2.35	.83	2.20	.61	6.78	1.75
Total	2.38	.83	2.36	.73	2.36	.67	7.10	1.77
12								
7th Grade	2.07	.78	1.93	.71	2.00	.66	6.00	1.73
13								
7th Grade	2.14	.83	2.18	.85	2.14	.71	6.45	2.04
8th Grade	2.08	.93	2.30	.93	2.19	.69	6.58	2.19
Total	2.10	.88	2.25	.89	2.17	.69	6.52	2.10
14								
7th Grade	2.67	.96	2.40	.83	2.73	.46	7.80	1.82
8th Grade	2.33	.87	2.78	.67	2.33	.87	7.44	1.94
Total	2.54	.93	2.54	.78	2.58	.65	7.67	1.83
15								
7th Grade	2.24	.90	2.36	.77	2.38	.82	6.99	2.13
16								
7th Grade	2.08	.79	2.42	.67	2.25	.62	6.75	1.60
17								
7th Grade	2.90	.94	3.05	.86	2.81	.81	8.76	2.32
18								
7th Grade	2.21	.87	2.21	.75	2.24	.82	6.67	2.09
8th Grade	2.59	.85	2.57	.90	2.63	.75	7.78	2.18
Total	2.42	.88	2.41	.85	2.45	.80	7.28	2.20
19								
7th Grade	3.06	.87	2.89	.76	2.94	.64	8.89	1.91
8th Grade	1.85	.76	1.80	.64	1.93	.69	5.59	1.69
Total	2.22	.97	2.14	.84	2.24	.82	6.59	2.32
20								
7th Grade	2.41	.70	2.38	.70	2.24	.65	7.03	1.78

Table 11

(Continued)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
21								
7th Grade	2.12	.80	1.94	.68	2.18	.60	6.24	1.70
8th Grade	2.46	.90	2.44	.72	2.34	.80	7.24	2.05
Total	2.31	.87	2.23	.74	2.27	.72	6.81	1.96
22								
8th Grade	3.17	.59	3.00	.87	3.20	.71	9.37	1.71
23								
8th Grade	2.78	.95	2.81	.84	2.78	.82	8.38	2.20
24								
7th Grade	3.29	.77	3.29	.69	3.29	.69	9.88	1.80
8th Grade	3.94	.24	3.82	.39	3.82	.39	11.59	.51
Total	3.62	.65	3.56	.61	3.56	.61	10.74	1.56
25								
7th Grade	2.17	.79	2.11	.68	2.00	.77	6.28	1.76
8th Grade	3.25	.79	3.08	.93	3.00	.72	9.33	2.08
Total	2.79	.95	2.67	.95	2.57	.89	8.02	2.46
26								
7th Grade	2.13	.76	2.16	.75	2.16	.72	6.44	1.76
8th Grade	2.40	.83	2.31	.74	2.29	.69	7.00	1.89
Total	2.25	.80	2.22	.75	2.22	.70	6.69	1.83
27								
8th Grade	2.06	.86	2.42	.77	2.28	.66	6.75	1.92
28								
7th Grade	2.27	.80	2.29	.78	2.18	.75	6.73	1.96
8th Grade	2.48	.85	2.47	.75	2.42	.74	7.37	1.90
Total	2.39	.83	2.40	.76	2.32	.75	7.11	1.94
29								
7th Grade	2.31	1.01	2.42	.64	2.25	.69	6.97	1.92
8th Grade	2.74	.81	2.74	.76	2.59	.80	8.07	2.04
Total	2.49	.95	2.56	.71	2.40	.75	7.44	2.03

Table 12

Distribution of Raw Scores Earned on Posttest by Experimental Pretest/Posttest Matched Pairs Subgroup Students by Reading, School and Grade (N=1,717)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1								
7th Grade	2.29	.81	2.58	.88	2.42	.72	7.29	2.14
8th Grade	2.74	.97	2.71	.86	2.71	.74	8.16	2.28
Total	2.55	.92	2.65	.87	2.58	.74	7.78	2.24
2								
7th Grade	1.90	.74	2.60	.70	2.40	.52	6.90	1.52
8th Grade	2.82	.85	2.73	.46	2.77	.61	8.32	1.43
Total	2.53	.92	2.69	.54	2.66	.60	7.88	1.58
3								
7th Grade	2.98	.79	2.86	.74	2.88	.73	8.73	1.86
8th Grade	2.90	.94	3.00	.82	3.04	.87	8.94	2.31
Total	2.95	.86	2.92	.79	2.95	.80	8.82	2.06
4								
7th Grade	2.54	.87	2.57	.87	2.41	.73	7.52	2.12
5								
7th Grade	2.27	.82	2.27	.87	2.16	.81	6.70	2.12
8th Grade	2.18	.80	2.06	.60	2.09	.51	6.32	1.30
Total	2.23	.80	2.18	.77	2.13	.69	6.54	1.81
6								
7th Grade	2.63	.81	2.81	.83	2.75	.58	8.19	1.72
8th Grade	3.00	.00	4.00	.00	3.00	.00	10.00	.00
Total	2.65	.79	2.88	.86	2.76	.56	8.29	1.72
7								
7th Grade	2.59	1.06	2.65	.70	2.35	.86	7.59	2.24
8th Grade	2.39	.84	2.45	.89	2.48	.72	7.32	2.01
Total	2.46	.92	2.52	.82	2.44	.77	7.42	2.07
8								
8th Grade	2.93	.73	3.07	.78	2.85	.66	8.85	1.73
9								
7th Grade	2.13	.64	2.25	.71	2.25	.46	6.63	1.30
8th Grade	2.43	.73	2.48	.79	2.48	.67	7.39	1.92
Total	2.35	.71	2.42	.76	2.42	.62	7.19	1.80

Table 12

(Continued)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
10								
7th Grade	1.98	.90	2.00	.77	2.06	.78	6.04	2.13
11								
7th Grade	2.82	.92	2.76	.90	2.67	.69	8.24	2.00
8th Grade	2.55	.81	2.60	.74	2.50	.64	7.65	1.78
Total	2.67	.87	2.67	.82	2.58	.67	7.92	1.89
12								
7th Grade	2.05	.73	2.12	.77	2.12	.77	6.29	1.99
13								
7th Grade	2.55	.86	2.77	.87	2.64	.66	7.95	1.15
8th Grade	2.65	.63	2.46	.65	2.35	.63	7.46	1.33
Total	2.60	.74	2.60	.76	2.48	.65	7.69	1.75
14								
7th Grade	2.40	.63	2.27	.59	2.33	.62	7.00	1.46
8th Grade	2.56	.88	2.11	1.05	2.22	.97	6.89	2.32
Total	2.46	.72	2.21	.78	2.29	.75	6.96	1.78
15								
7th Grade	2.43	.94	2.39	.89	2.28	.77	7.11	2.17
16								
7th Grade	2.00	.85	2.33	.78	2.42	.51	6.75	1.82
17								
7th Grade	3.38	.50	2.90	.70	2.67	.80	8.95	1.63
18								
7th Grade	2.29	.94	2.07	.84	2.12	.71	6.48	2.10
8th Grade	2.35	.87	2.49	.86	2.59	.88	7.43	2.28
Total	2.32	.90	2.30	.87	2.38	.83	7.00	2.24
19								
7th Grade	3.39	.70	3.00	.69	3.11	.76	9.50	1.79
8th Grade	2.15	.76	2.15	.76	2.15	.76	6.44	2.28
Total	2.53	.94	2.41	.83	2.44	.88	7.37	2.56
20								
7th Grade	2.85	.74	2.56	.61	2.50	.51	7.91	1.26

Table 12

(Continued)

School	First Reading		Second Reading		Third Reading		Total Reading	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
21								
7th Grade	2.46	.89	2.40	.83	2.34	.75	7.20	2.03
8th Grade	2.76	.81	2.74	.75	2.79	.78	8.29	1.95
Total	2.64	.85	2.59	.80	2.60	.80	7.83	2.05
22								
8th Grade	3.50	.68	3.50	.68	3.50	.68	10.50	2.05
23								
8th Grade	2.86	.92	3.00	.82	2.68	.75	8.54	2.19
24								
7th Grade	3.65	.61	3.47	.51	3.47	.62	10.59	1.46
8th Grade	3.94	.24	3.88	.33	3.76	.44	11.59	.51
Total	3.79	.48	3.68	.47	3.62	.55	11.09	1.19
25								
7th Grade	1.89	.68	1.94	.80	2.06	.80	5.89	1.91
8th Grade	3.21	.78	3.13	.61	3.25	.61	9.58	1.38
Total	2.64	.98	2.62	.91	2.74	.91	8.00	2.45
26								
7th Grade	2.36	.76	2.49	.76	2.47	.70	7.31	1.79
8th Grade	2.60	.87	2.56	.88	2.51	.84	7.67	2.25
Total	2.46	.82	2.52	.81	2.49	.76	7.47	2.01
27								
8th Grade	2.39	.84	2.42	.87	2.42	.77	7.22	2.10
28								
7th Grade	2.30	.90	2.32	.77	2.34	.80	6.96	2.09
8th Grade	2.54	.83	2.54	.67	2.51	.71	7.58	1.74
Total	2.44	.87	2.45	.72	2.44	.75	7.33	1.91
29								
7th Grade	2.22	.80	2.36	.76	2.22	.68	6.81	1.88
8th Grade	2.67	.83	2.78	.75	2.52	.64	7.96	1.81
Total	2.41	.84	2.54	.78	2.35	.68	7.30	1.92

Table 13

Distribution of Raw Scores Earned by Comparison and Experimental Group Students by Reading Across Schools and Grades (N=4,071)

Group	Pretest								Posttest							
	Reading 1		Reading 2		Reading 3		Reading Total		Reading 1		Reading 2		Reading 3		Reading Total	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Comparison	--	--	--	--	--	--	--	--	2.24	.82	2.29	.76	2.32	.74	6.85	1.93
Experimental																
Pretest Only	2.31	.90	2.33	.82	2.31	.79	6.94	2.15	--	--	--	--	--	--	--	--
Posttest Only	--	--	--	--	--	--	--	--	2.53	.92	2.54	.87	2.46	.84	7.53	2.29
Pretest/Posttest Matched Pairs	2.37	.89	2.37	.82	2.34	.78	7.08	2.12	2.56	.90	2.56	.84	2.52	.79	7.64	2.18

Table 14

Comparison Group Analysis of Variance:
Posttest Total Score by School
(N=648)

Source of Variation	df	Sum of Squares	Mean Squares	F	P
Between Groups	7	348.36	49.77	15.48	.001
Within Groups	640	2058.10	3.22		
Total	647	2406.46			

Table 15

Experimental Pretest Only Subgroup Analysis of
Variance: Pretest Total by School
(N=804)

Source of Variation	df	Sum of Squares	Mean Squares	F	P
Between Groups	30	756.59	25.22	6.59	.001
Within Groups	773	2956.77	3.83		
Total	803	3713.36			

Table 16

Experimental Posttest Only Subgroup Analysis of Variance: Posttest Total by School (N=902)

Source of Variation	df	Sum of Squares	Mean Squares	F	P
Between Groups	28	987.71	35.28	8.29	.001
Within Groups	873	3717.02	4.26		
Total	901	4704.72			

Table 17

Experimental Pretest/Posttest Matched Pairs Subgroup Analysis of Variance: Pretest Total by School (N=1,717)

Source of Variation	df	Sum of Squares	Mean Squares	F	P
Between Groups	28	1121.24	40.04	10.21	.001
Within Groups	1688	6621.93	3.92		
Total	1716	7743.17			

Table 18

Experimental Pretest/Posttest Matched Pairs
Subgroup Analysis of Variance: Posttest
Total by School
(N=1,717)

Source of Variation	df	Sum of Squares	Mean Squares	F	P
Between Groups	28	1385.55	49.48	12.30	.001
Within Groups	1688	6792.23	4.02		
Total	1716	8177.78			

The Theoretical Model of Holistic Essay Scoring

Theoretically, the scores for a large-scale holistic essay reading should approximate the normal curve. This is due primarily to the fact that the essays for this type of scoring are judged in relation to other papers in the population; they are not judged against a preconceived ideal. Therefore, one would expect that within one reading the number of papers given scores of 'two' and 'three' would be higher than those given scores of 'one' and 'four.' One would also expect that the number of papers given scores of 'one' and 'four' would be similar as would the number of papers given scores of 'two' and 'three.'

Statistics calculated to determine the degree to which a distribution of cases approximates a normal curve include the third and fourth moments; skewness and kurtosis, respectively. Skewness will equal zero when the distribution represents a bell-shaped curve which is completely symmetrical. Kurtosis will also equal zero when the distribution repre-

sents a perfect normal curve. The kurtosis and skewness statistics were used to test the following hypothesis:

H_1: The distribution of scores for each of the three readings and summed totals across pretest, posttest, seventh- and eighth-grades, comparison and experimental groups, and all schools, should approximate a normal distribution.

Table 19 shows the distribution of cases for each reading of all papers across all groups including obtained measures of skewness and kurtosis. The positive values of skewness indicate the deviations from perfect symmetry are such that there is a clustering to the left of the mean with extreme values to the right. Negative kurtosis values, across all readings, show the distributions to be flatter than a normal distribution.

Table 19

Distribution of Raw Scores by Reading
(N=5,788)

Score	First Reading		Second Reading		Third Reading	
	N	%	N	%	N	%
1	854	14.8	664	11.5	615	10.6
2	2,381	41.1	2,541	43.9	2,733	47.2
3	1,781	30.8	1,963	33.9	1,919	33.2
4	772	13.3	620	10.7	521	9.0
Mean	2.43		2.44		2.41	
Standard Deviation	.90		.83		.80	
Skewness	.16		.16		.22	
Kurtosis	-.73		-.53		-.38	

In terms of skewness, the first and second readings appear to best approximate the normal curve. The third reading, on the other hand, appears to best approximate the normal curve in regard to kurtosis. While specific tests of significance were not performed, the statistics indicate that the null hypothesis that the distribution of cases for each of the three readings does not approximate the normal curve, can be rejected with some certainty.

Values of skewness and kurtosis were also calculated for the distribution of total scores summed across the three readings. The range of total scores was three to twelve. This distribution as presented in Table 20, also indicates scores to cluster to the left of the mean with extreme values to the right, and that the distribution is flatter than that of a perfectly normal distribution; skewness of .22 and kurtosis of -.49.

Table 20

Distribution of Total Raw Scores
(N=5,788)

Score	N	Percent
3	208	3.6
4	334	5.8
5	640	11.1
6	1,145	19.8
7	973	16.8
8	793	13.7
9	777	13.4
10	414	7.2
11	282	4.9
12	222	3.8
Mean	7.27	
Standard Deviation	2.17	
Kurtosis	-.49	
Skewness	.22	

Results indicated that the null hypothesis that the distribution of cases for the summed total scores does not approximate the normal curve, could be rejected with some certainty.

Interrater Reliability

The value of scores is dependent upon their objectivity and reliability. To the degree that qualified observers would assign different scores to the same paper, the measurement lacks objectivity and utility. Before proceeding to analyses dependent upon assigned scores, it was necessary to establish the interrater reliabilities and to test the following hypotheses:

H_{2a}: The interrater reliabilities reflect consistency in scoring.

H_{2b}: The number of papers with discrepant scores is less than the number of papers with non-discrepant scores.

H_{2c}: The number of papers receiving the same score by each of the three readers (perfect agreement) is greater than the number of papers not reflecting perfect agreement.

Table 21 presents Cronbach's alpha coefficients of reliability for the reading of all student essays by study subgroups. The reliability coefficients ranged from .77 to .84. Discrepant scores were evident in 258 of the 5,788 papers read and scored (4.46 percent as reported in Chapter III). The agreement and disagreement among the three readings for the various population subgroups is presented in Tables 59 through 78 in Appendix D. In only one crosstabulation are the number of papers receiving the same score by each of the three readers less than the number of papers not reflecting perfect agreement.

Clearly, the reading reliability was not so strong so as to suggest perfect agreement between readers in scoring. However, the null hypotheses that the interrater reliabilities did not reflect consistency in scoring, that the number of papers with discrepant scores was equal to or greater than the number of papers with non-discrepant scores, and that the number of papers indicating perfect agreement among the three scores was equal to or less than those not reflecting perfect agreement, could be rejected with confidence. The obtained reliability coefficients seemed substantial enough to proceed with the other study analyses.

Table 21

Alpha Coefficients of Reliability by Group and Grade

Group	Alpha
All experimental group pretests	
7th Grade	.80
8th Grade	.83
All experimental group posttests	
7th Grade	.82
8th Grade	.84
All comparison group posttests	
7th Grade	.77
8th Grade	.78
All posttests (Comparison and Experimental Groups)	
7th Grade	.81
8th Grade	.83

Differences Between Seventh- and Eighth-Graders

Various t-tests comparing sample total score means were performed by population subgroups to determine if differences existed between seventh- and eighth-grade students on the pretest and posttest in testing the following hypothesis:

H_3: Eighth-grade students will show higher total essay scores than seventh-grade students within each of the pretest and posttest administrations.

The t-tests included those for independent samples where cases were classified into two groups and a test of mean differences was performed. Such tests of significance were performed for the two distinct groups; pretest and posttest (experimental and comparison groups).

The experimental group was comprised of two subgroups for the pretest: students having taken only the pretest (pretest only subgroup); and students having taken both the pretest and posttest (matched pairs). The two subgroups could not be combined by pretest unless statistically, there was no difference between the pretest only seventh-graders and the pretest/posttest matched pairs seventh-graders, and likewise between the pretest only and pretest/posttest matched pairs eighth-graders. If differences within grade by subgroup were shown to exist, then comparisons between the grade levels could be analyzed only individually by subgroup. For all t-tests between independent groups, an approximation to t was computed based on separate variances where sample variances were unequal. t, based on the pooled variances was computed where sample variances were not unequal as indicated by the F-test of sample variances.

Tables 22 and 23 present one-tailed t-test results to determine

differences on the pretest between the seventh- and eighth-grades. The stated null hypothesis was that the seventh-grade mean was equal to or greater than the eighth-grade mean.

Table 22

t-Test Between Seventh- and Eighth-Grade Pretest Means for the Experimental Pretest/Posttest Matched Pairs Subgroup (N=1,717)

Group	N	Mean	t	P
7th Grade	894	6.84	*4.81	.000
8th Grade	823	7.34		

*Approximation to t based on separate variances.

Table 23

t-Test Between Seventh- and Eighth-Grade Pretest Means for the Experimental Pretest Only Subgroup (N=804)

Group	N	Mean	t	P
7th Grade	437	6.58	*5.28	.000
8th Grade	367	7.37		

*Approximation to t based on separate variances.

The null hypothesis that the seventh-grade mean is equal to or greater than the eighth-grade mean for each of the two experimental subgroups having taken the pretest was rejected beyond the .001 significance level.

To determine whether differences existed between the seventh- and eighth-grade total score means for both pretest experimental groups combined, a two-tailed t-test was first performed between the seventh-graders of the pretest only subgroup and the pretest/posttest matched pairs subgroup. The same test was performed between the means of both groups for the eighth-graders. Results are shown in Tables 24 and 25.

Table 24

t-Test of Seventh-Grade Pretest Means Between the Pretest Only and Pretest/Posttest Matched Pairs Experimental Subgroups (N=1,331)

Group	N	Mean	t	P
Pretest Only	437	6.58	*2.26	.024
Pretest/Posttest Matched Pairs	894	6.84		

*t based on pooled variances.

Table 25

t-Test of Eighth-Grade Pretest Means Between the Pretest Only and Pretest/Posttest Matched Pairs Experimental Subgroups (N=1,190)

Group	N	Mean	t	P
Pretest Only	367	7.37	* .27	.788
Pretest/Posttest Matched Pairs	823	7.34		

*t based on pooled variances.

The null hypothesis that there was a difference between the seventh-grade means of the pretest only and pretest/posttest matched pairs experimental subgroups was not rejected. The null hypothesis that there was a difference between the eighth-grade means of the pretest only and the pretest/posttest matched pairs experimental subgroups was rejected.

Since the seventh-grade means across both pretest experimental subgroups were statistically different, combining the seventh-graders from the pretest only and pretest/posttest matched pairs subgroups would lead to misleading results. Therefore, eighth-graders were also not combined across experimental subgroups. In summary, however, the eighth-grade mean was higher than the seventh-grade mean for both pretest experimental groups as would be expected.

Tables 26 to 28 present one-tailed t-test results to determine differences on the posttest between the seventh- and eighth-graders; the stated null hypothesis that the seventh-grade mean is equal to or greater than the eighth-grade mean for each posttest subgroup.

Table 26

t-Test Between Seventh- and Eighth-Grade Posttest Means for the Comparison Group (N=648)

Group	N	Mean	t	P
7th Grade	179	6.78	*.54	.295
8th Grade	469	6.87		

*t based on pooled variances.

For the comparison group, the null hypothesis that the seventh-grade posttest mean is equal to or greater than the eighth-grade posttest mean was not rejected.

Table 27

t-Test Between Seventh- and Eighth-Grade Posttest Means for the Experimental Pretest/Posttest Matched Pairs Subgroup (N=1,717)

Group	N	Mean	t	P
7th Grade	894	7.35	*5.83	.000
8th Grade	823	7.96		

*t based on pooled variances.

For the experimental pretest/posttest matched pairs subgroup, the null hypothesis that the seventh-grade posttest mean is equal to or greater than the eighth-grade posttest mean was rejected beyond the .001 level of significance.

Table 28

t-Test Between Seventh- and Eighth-Grade Posttest Means for the Experimental Posttest Only Subgroup (N=902)

Group	N	Mean	t	P
7th Grade	505	7.48	* .78	.437
8th Grade	397	7.60		

*t based on pooled variances.

For the experimental posttest only subgroup, the null hypothesis that the seventh-grade posttest mean is equal to or greater than the eighth-grade mean was not rejected.

Two-tailed t-tests were performed between the posttest seventh-grade means of the posttest only subgroup and the pretest/posttest matched pairs subgroup, and also for the posttest eighth-grade means of both experimental posttest groups. This was done in order to determine if all experimental seventh-grade scores could be compared to all experimental eighth-grade scores. Results of the two t-tests are shown in Tables 29 and 30.

Table 29

t-Test of Seventh-Grade Posttest Means Between the Posttest Only and Pretest/Posttest Matched Pairs Experimental Subgroups (N=1,399)

Group	N	Mean	t	P
Posttest Only	505	7.48	*1.06	.290
Pretest/Posttest Matched Pairs	894	7.35		

*t based on pooled variances.

While there was no difference for the seventh-grade posttest means between the two experimental posttest subgroups, a significant difference for the eighth-grade posttest means of the two subgroups was noticed. Therefore, combining across subgroups to test the difference between all experimental seventh-graders and all eighth-graders on the posttest would provide misleading results.

Table 30

t-Test of Eighth-Grade Posttest Means Between the Posttest Only and Pretest/Posttest Matched Pairs Experimental Subgroups (N=1,291)

Group	N	Mean	t	P
Posttest Only	397	7.39	*5.15	.000
Pretest/Posttest Matched Pairs	894	7.84		

*t based on pooled variances.

Growth in Writing Ability Over Time

Analyses to determine whether or not there was growth in writing ability over time were performed to test the following hypothesis:

H_4: Obtained pretest and posttest total scores reflect growth in writing ability over time within and across grade levels.

These analyses included only the pretest/posttest matched pairs experimental subgroup. This was due partially because in fact, there were differences in the pretest total scores between the pretest only and the pretest/posttest matched pairs subgroups, as well as differences in the posttest total scores between the posttest only and the pretest/posttest matched pairs subgroups. Additionally, the use of matched pairs receives attention in the literature.

The minimum total score for both pretest and posttest was three; the maximum total score for each essay administration was twelve. The maximum possible loss from pretest to posttest was nine points; the

maximum possible gain from pretest to posttest was also nine points. Analyses to determine gain, loss, or no change from pretest to posttest were performed for each of grades seven and eight, as well as for both grades combined.

Table 31 shows the number of students and corresponding percentages of students who performed better on the posttest than on the pretest, who exhibited no change in total score from pretest to posttest, and who performed better on the pretest than the posttest for the seventh-grade, eighth-grade, and combined groups.

More students in both grades seven and eight performed better on the posttest than on the pretest than students either exhibiting no change or a negative gain. Students gaining comprised almost fifty percent of the group.

Table 32 presents the pretest to posttest change values for both grades individually and combined. The mean change value for seventh-graders was .50 with a standard deviation of 2.0. The mean change value for eighth-graders was .21 with a standard deviation of .87. The mean change value for the total group was .56 with a standard deviation of 2.1.

It must be noted that due to the regression effect, or regression towards the mean, students with the lowest pretest scores can appear to gain more than students with higher initial scores. Raw change or gain scores formed by subtracting pretest scores from posttest scores, as shown in Table 32 can lead to fallacious conclusions, because these scores are systematically related to random error of measurement. The mean gain scores as reported above do not seem substantial. A different

Table 31

Distribution of Pretest to Posttest Gain Scores
(N=1,717)

Group	Total N	Positive Gain N	Positive Gain %	No Gain N	No Gain %	Negative Gain N	Negative Gain %
7th Grade	894	444	49.66	174	19.46	276	30.87
8th Grade	823	413	50.18	170	20.66	240	29.16
Total	1,717	857	49.91	344	20.03	517	30.00

Table 32

Distribution of Pretest to Posttest Change Values

Change Value	Group 7th Grade N	7th Grade %	8th Grade N	8th Grade %	Total N	Total %
+9	--	--	1	.1	1	.1
+8	--	--	--	--	--	--
+7	1	.1	3	.4	4	.2
+6	8	.9	11	1.3	19	1.1
+5	11	1.2	19	2.3	30	1.7
+4	44	4.9	41	5.0	85	5.0
+3	72	8.1	72	8.7	144	8.4
+2	137	15.3	126	15.3	263	15.3
+1	171	19.1	140	17.0	311	18.1
0	174	19.5	170	20.6	344	20.0
-1	131	14.7	116	14.2	247	14.4
-2	85	9.5	69	8.4	154	9.0
-3	46	5.2	37	4.5	83	4.8
-4	10	1.1	9	1.1	19	1.1
-5	2	.2	7	.8	9	.5
-6	1	.1	1	.1	2	.1
-7	--	--	--	--	--	--
-8	--	--	1	.1	1	.1

picture of growth in writing ability over time, however, is apparent when significance tests are performed to determine if the posttest means are reliably different from and greater than the pretest means. The difference in sample means for pretest and posttest is a better estimate of the mean difference than is the analysis of raw gain or change scores.

Correlated t-tests for paired samples were performed for seventh-grade, eighth-grade, and combined grades between pretest and posttest means. Results are shown in Tables 33 to 35.

Table 33

Correlated t-Test Between Pretest and Posttest Means for Seventh-Graders (N=894)

Group	Mean	t	P
Pretest	6.84	7.51	.000
Posttest	7.35		

The null hypothesis that there is no difference between pretest and posttest means was rejected beyond the .001 significance level for seventh- and eighth-grades individually and combined.

Table 34

Correlated t-Test Between Pretest and Posttest Means for Eighth-Graders (N=823)

Group	Mean	t	P
Pretest	7.34	8.22	.000
Posttest	7.96		

Table 35

Correlated t-Test Between Pretest and Posttest Means for all Matched Pair Students (N=1,717)

Group	Mean	t	P
Pretest	7.08	11.13	.000
Posttest	7.64		

The Effectiveness of the Writer's Clinic Program

Comparison Versus Experimental Groups

Posttest differences between the experimental (treatment) and comparison (no treatment) groups were analyzed to test the following hypothesis:

H_5: The obtained total posttest scores of experimental students will be greater than those of comparison students within

and across grade levels.

Before these differences were analyzed, however, it was necessary to determine whether or not the two groups were essentially equivalent at pretest time. The comparison group was selected after experimental students had taken the pretest. Therefore, the comparison group students were administered the posttest only. Without the pretest for comparison students, a t-test between experimental and comparison means was not possible.

In order to compare the status of program participants following treatment to that of similar non-participants, some type of reference group is critical. Members of the reference group should be as much like those whose teachers participated in the program with respect to variables such as age, grade, sex, class, and school. While a one-to-one match is not necessary, the overall group profiles should be similar.

A comparison group is one in which students similar to students whose teachers participated in the program are identified. They are not necessarily randomly assigned as in control groups. Nevertheless, comparison students and experimental students should be administered the same instruments on the same schedule. One limitation of this study however, was that comparison students did not take the pretest. While the comparison group was formulated so as to match the characteristics of 700 randomly selected experimental students, the matching and hence, representativeness of the comparison group was less than ideal. For example, the proportions of seventh- and eighth-graders in each group seemed rather dissimilar; 27.6 percent of comparison students were seventh-graders as compared to 53.6 percent seventh-graders in the experimental group. The

experimental group was relatively balanced with respect to grade, whereas the comparison group was comprised of proportionately more eighth-graders. Additionally, the comparison group students came from eight schools. Experimental group students were enrolled at only five of the eight comparison group schools. Therefore, there were students in the comparison group from three schools having no experimental students. Based on these descriptive data, equivalency of comparison group students and experimental group students at pretest time was tenuous.

One test that might have indicated some equivalency between groups at pretest time was the comparison of the seventh-grade comparison group posttest mean with the eighth-grade experimental group pretest mean. Presumably, end-of-year seventh-graders would be similar in performance to beginning-of-year eighth-graders. Three two-tailed t-tests between means were performed. Tables 36 to 38 present the findings.

Table 36

t-Test Between Means of All Seventh-Grade Comparison Group Posttests and Eighth-Grade Experimental Pretest Only Subgroup (N=546)

Group	N	Mean	t	P
7th Grade Comparison Posttest Only Group	179	6.78	*3.18	.001
8th Grade Experimental Pretest Only Group	367	7.37		

*Approximation to t based on separate variances.

Table 37

t-Test Between Means of All Seventh-Grade Comparison Group Posttests and Eighth-Grade Experimental Pretest/Posttest Matched Pairs Pretest (N=1,002)

Group	N	Mean	t	P
7th Grade Comparison Posttest Only Group	179	6.78	*3.38	.001
8th Grade Experimental Matched Pairs Group	823	7.34		

*Approximation to t based on separate variances.

Table 38

t-Test Between Means of All Seventh-Grade Comparison Group Posttests and All Eighth-Grade Experimental Pretests (N=1,369)

Group	N	Mean	t	P
7th Grade Comparison Posttest Only Group	179	6.78	*3.57	.001
8th Grade Experimental Pretest Groups	1,190	7.35		

*Approximation to t based on separate variances.

The null hypothesis that there is a difference between the seventh-grade comparison group posttest mean and the eighth-grade experimental group pretest mean was not rejected. Equivalency of the comparison and experimental groups at pretest time could not be established. It should

be noted that experimental students took the pretest in December; not quite the beginning of the school year. Had the pretest been administered to experimental students in September, the t-tests performed might have indicated no difference between groups.

Keeping the design limitations in mind and not establishing equivalency of comparison and experimental groups at pretest time, comparisons between the comparison group and experimental group posttests were nevertheless, conducted. Tables 39 through 47 present results of these comparisons.

The null hypothesis that the experimental group posttest mean was less than or equal to the comparison group posttest mean was rejected beyond the .001 significance level for each grade individually and combined and for each experimental posttest subgroup. Such significant differences between comparison and experimental group posttest means would signify a true difference between the groups and would indicate Writer's Clinic program effectiveness. The uncertainty however, as to whether or not the groups were equivalent at pretest time puts such interpretations into jeopardy.

Table 39

t-Test Between Seventh-Grade Posttest Means of Comparison Group and Experimental Pretest/Posttest Matched Pairs Subgroup (N=1,073)

Group	N	Mean	t	P
Comparison Group	179	6.78	*3.27	.001
Experimental Pretest/ Posttest Matched Pairs	894	7.35		

*t based on pooled variances.

Table 40

t-Test Between Seventh-Grade Posttest Means of Comparison Group and Experimental Posttest Only Subgroups (N=684)

Group	N	Mean	t	P
Comparison Group	179	6.78	*3.94	.000
Experimental Posttest Only	505	7.48		

*Approximation to t based on separate variances.

Table 41

t-Test Between Seventh-Grade Posttest Means of Comparison and Experimental Groups (N=1,578)

Group	N	Mean	t	P
Comparison Group	179	7.78	*3.92	.000
Experimental Group	1,399	7.39		

*Approximation to t based on separate variances.

Table 42

t-Test Between Eighth-Grade Posttest Means of Comparison Group and Experimental Pretest/Posttest Matched Pairs Subgroup (N=1,292)

Group	N	Mean	t	P
Comparison Group	469	6.88	*9.23	.000
Experimental Pretest/ Posttest Matched Pairs	823	7.96		

*Approximation to t based on separate variances.

Table 43

t-Test Between Eighth-Grade Posttest Means of Comparison Group and Experimental Posttest Only Subgroup (N=866)

Group	N	Mean	t	P
Comparison Group	469	6.88	*4.90	.000
Experimental Posttest Only	397	7.60		

*Approximation to t based on separate variances.

Table 44

t-Test Between Eighth-Grade Posttest Means of Comparison and Experimental Groups (N=1,689)

Group	N	Mean	t	P
Comparison Group	469	6.88	*8.78	.000
Experimental Group	1,220	7.84		

*Approximation to t based on separate variances.

Table 45

t-Test Between Posttest Means of Comparison Group and Experimental Pretest/Posttest Matched Pairs Subgroup Across Grades (N=2,365)

Group	N	Mean	t	P
Comparison Group	648	6.85	*8.55	.000
Experimental Pretest/ Posttest Matched Pairs	1,717	7.64		

*Approximation to t based on separate variances.

Table 46

t-Test Between Posttest Means of Comparison Group and Experimental Posttest Only Subgroup Across Grades (N=1,550)

Group	N	Mean	t	P
Comparison Group	648	6.85	*6.32	.000
Experimental Posttest Only	902	7.53		

*Approximation to t based on separate variances.

Table 47

t-Test Between Posttest Means of Comparison and Experimental Groups Across Grades (N=3,267)

Group	N	Mean	t	P
Comparison Group	648	6.85	*8.60	.000
Experimental Group	2,619	7.60		

*Approximation to t based on separate variances.

The Relationship Between Holistic Essay Scores and Classroom Composition Grades and Teacher Ratings of Overall Writing Ability

The validity of an assessment instrument refers to the extent to which the instrument measures what it is intended to measure. This study attempted to determine the relationship between holistic essay scores and two contemporary criteria for writing: grades assigned to compositions written by students for English course requirements; and teacher ratings of overall student writing ability. The criteria data were collected to test the following hypotheses:

H_{6a}: There is a relationship between grades assigned to compositions at the beginning of the year and pretest holistic essay scores.

H_{6b}: There is a relationship between grades assigned to compositions at the end of the year and posttest holistic essay scores.

H_{6c}: There is a relationship between teacher ratings of students' overall writing ability at the beginning of the year and pretest holistic essay scores.

H_{6d}: There is a relationship between teacher ratings of students' overall writing ability at the end of the year and posttest holistic essay scores.

Teachers supplied composition grades for 215 students who had taken both the pretest and posttest for the Writer's Clinic evaluation. The grades were reported for the months of December through May. Additionally, teachers rated 173 of the pretest/posttest matched pairs

students on overall writing ability both at the beginning of the year and at the end of the year on a scale of one to four (four was the highest rating).

Composition grades were split into two groups: those assigned in December, January and February were labeled as pretest essay grades: those assigned in March, April and May were labeled as posttest essay grades. Before determining whether or not there was a relationship between composition grades and essay test scores, t-tests between the pretest composition grades and posttest composition grades were run to determine whether or not student composition grades were higher towards the end of the school term. Tables 48 through 50 show the t-test results for seventh-graders, eighth-graders and both grades combined, respectively.

Table 48

Correlated t-Test Between Means of Pretest and Posttest Composition Grades for Seventh-Grade Students (N=135)

Group	Mean	T	P
Pretest Composition Grades	2.90	4.58	.000
Posttest Composition Grades	3.07		

Table 49

Correlated t-Test Between Means of Pretest and Posttest Composition Grades for Eighth-Grade Students (N=80)

Group	Mean	t	P
Pretest Composition Grades	2.95	.62	.536
Posttest Composition Grades	2.98		

Table 50

Correlated t-Test Between Means of Pretest and Posttest Composition Grades Across Grade Levels (N=215)

Group	Mean	t	P
Pretest Composition Grades	2.92	4.09	.000
Posttest Composition Grades	3.04		

While there was not a significant difference between pre- and posttest composition grades for eighth-grade students, the null hypothesis of no difference across grades was rejected beyond the .001 level of significance. Student essay grades were therefore, higher in general, at the end of the year than at the beginning of the year.

Correlated t-tests were also performed to determine whether or not a difference existed between teacher ratings of student writing ability at the beginning of the year and at the end of the year. These data are

presented in Tables 51 through 53.

Table 51

Correlated t-Test Between Means of Pretest and Posttest Teacher Ratings of Overall Student Writing Ability for Seventh-Grade Students (N=116)

Group	Mean	t	P
December Ratings	2.51	7.13	.000
May Ratings	2.96		

Table 52

Correlated t-Test Between Means of Pretest and Posttest Teacher Ratings of Overall Student Writing Ability for Eighth-Grade Students (N=57)

Group	Mean	t	P
December Ratings	2.07	8.55	.000
May Ratings	2.84		

Table 53

Correlated t-Test Between Means of Pretest and Posttest Teacher Ratings of Overall Student Writing Ability Across Grade Levels (N=173)

Group	Mean	t	P
December Ratings	2.36	10.52	.000
May Ratings	2.92		

Clearly, teachers' ratings of student overall writing ability reflected better writing at the end of the year than in December. This indicates growth in student writing ability over time for seventh- and eighth-graders alike. The null hypothesis of no difference was rejected beyond the .001 significance level for pretest to posttest ratings within and across the grade levels.

As would be expected, pretest and posttest scores for the students for whom the validity criteria were supplied were significantly different; higher scores at posttest time. Tables 54 through 56 present the t-test results for the test variables for seventh-, eighth-, and combined grade levels, respectively.

As was expected, posttest scores were significantly higher than pretest scores for both the seventh- and eighth-graders. The null hypothesis of no difference was rejected beyond the .01 level for seventh-graders and beyond the .001 level of significance for eighth-graders and both grades combined.

Table 54

Correlated t-Test Between Means of Seventh-Grade Pre- and Posttest Scores (N=135)

Group	Mean	t	P
Pretest	7.34	2.74	.007
Posttest	7.80		

Table 55

Correlated t-Test Between Means of Eighth-Grade Pre- and Posttest Scores (N=80)

Group	Mean	t	P
Pretest	6.96	4.43	.000
Posttest	8.05		

Table 56

Correlated t-Test Between Means of Pretest and Posttest Scores Across Grades (N=215)

Group	Mean	t	P
Pretest	7.20	4.93	.000
Posttest	7.89		

In summary, these analyses showed teacher ratings, composition grades, and essay test scores to reflect growth in writing over time.

Pearson correlation coefficients were calculated to determine the relationships between composition grades and test scores and teacher ratings and test scores. Table 57 shows the means of each of the six validity component variables by grade. Correlation coefficients are presented in Table 58.

Table 57

Group Means by Validity Variables and by Grade

Variable	Scale		Grade		
	Low	High	7th	8th	7th & 8th
Pretest Compositions	1	5	2.90	2.95	2.92
Posttest Compositions	1	5	3.07	2.98	3.04
Teacher Ratings December	1	4	2.51	2.07	2.36
Teacher Ratings May	1	4	2.96	2.84	2.91
Pretest Essay Scores	3	12	7.34	6.96	7.20
Posttest Essay Scores	3	12	7.80	8.05	7.89

Table 58

Pearson Correlation Coefficients

Criterion Variables	7th Grade Pretest	7th Grade Posttest	8th Grade Pretest	8th Grade Posttest	Across Grades Pretest	Across Grades Posttest
Teacher Ratings December	.25***	--	.15	--	.26***	--
Teacher Ratings May	--	.55***	--	.49***	--	.53***
Pretest Composition Grades	.03	--	.43***	--	.16**	--
Posttest Composition Grades	--	.02	--	.34***	--	.13*

Significance Levels: *p<.05
**p<.01
***p<.001

The Pearson correlation coefficients indicate there are relationships between holistic essay test scores and teacher ratings of overall writing ability and composition grades. The strongest relationship lies in the correlation between teacher ratings and holistic essay scores; .26 for December ratings and .53 for May ratings. The lower correlation with December ratings rests with the fact that teachers rated the students in May (they were asked to recall student writing ability as it was five months earlier and as it was in May). It seems reasonable to assume that composition grades were referred to in order to rate the students' ability as it was at pretest time. The lower correlation between holistic essay test scores and composition grades is therefore, reflected in the December ratings. Lower correlations of test scores with grades also seems likely due to the fact that grades are distinctly

different from scores; they take into account individual abilities, the topic responded to and the structure of the classroom curriculum. Still, the criteria of both composition grades and teacher ratings were related to obtained holistic essay scores and therefore, serve as acceptable predictors of concurrent validity to some extent.

Summary of Findings

The distribution of test scores for each of the three readings and summed totals approximated the normal distribution; the theoretical model of holistic essay scoring was supported by the case study. Tests of significance were not performed to test the following hypothesis:

H_1: The distribution of scores for each of the three readings and summed totals across pretest, posttest, seventh- and eighth-grades, comparison and experimental groups, and all schools, should approximate a normal distribution.

However, the null hypothesis that the distribution of scores did not approximate the normal distribution was rejected with some certainty by examining the score frequency distributions for the readings individually and combined and by examining the measures of skewness and kurtosis.

Consistency in scoring was analyzed by testing three hypotheses:

H_{2a}: The interrater reliabilities reflect consistency in scoring.

H_{2b}: The number of papers with discrepant scores is less than the number of papers with non-discrepant scores.

H_{2c}: The number of papers receiving the same score by each of the three readers (perfect agreement) is greater than the number of papers not reflecting perfect agreement.

Relatively strong agreement between readers was substantiated by reader reliability coefficients ranging from .77 to .84. Discrepancies in scoring among the three readers per essay paper were found in only 4.46 percent of the 5,788 papers included for analyses. Additionally, the number of papers reflecting perfect agreement was greater than the number of papers not reflecting perfect agreement. This reading reliability increased confidence in the validity of the assigned scores and enabled further analyses of the data.

It was hypothesized that students at a higher educational grade level would perform better than students at a lower educational level in testing the following hypothesis:

H_3: Eighth-grade students will show higher total essay scores than seventh-grade students within each of the pretest and posttest administrations.

Within each of the pretest and posttest essay groups, eighth-grade students consistently performed better than seventh-graders in terms of total essay scores. The null hypothesis of no difference in scores between the grade levels was rejected beyond the .05 level of significance.

Growth in writing ability over time was tested by one hypothesis:

H_4: Obtained pretest and posttest total scores reflect growth in writing ability over time within and across grade levels.

Growth in writing ability over a close to five-month period was substantiated. The null hypothesis of no difference between pretest and posttest mean scores within the experimental group was rejected beyond the .001 significance level for seventh- and eighth-graders individually and combined.

Writer's Clinic program effectiveness was tested by the following hypothesis:

H_5: The obtained total posttest scores of experimental students will be greater than those of comparison students within and across grade levels.

Determining the effectiveness of the teacher in-service program on writing was difficult to establish primarily because comparison (no treatment) students did not take the pretest and it could not be said with any degree of certainty that this group was similar to the experimental (treatment) group at pretest time. Alternate analyses conducted to try to establish equivalency of comparison and experimental group students at pretest time were not successful. Keeping the design limitations in mind, posttest comparisons made between the treatment and no treatment groups indicated a difference significant beyond the .001 level of significance; experimental students' posttest scores being higher than comparison group students' scores.

The study also attempted to determine the relationship of holistic essay scores with student class composition grades and teacher ratings of student writing ability by testing four hypotheses:

H_{6a}: There is a relationship between grades assigned to compositions at the beginning of the year and pretest holistic essay scores.

H_{6b}: There is a relationship between grades assigned to compositions at the end of the year and posttest holistic essay scores.

H_{6c}: There is a relationship between teacher ratings of students'

overall writing ability at the beginning of the year and pretest holistic essay scores.

H_{6d}: There is a relationship between teacher ratings of students' overall writing ability at the end of the year and posttest holistic essay scores.

Relationships were established for each of the four hypotheses by Pearson correlation coefficients. The strongest relationship was between posttest holistic essay scores and teacher ratings of student writing ability in May (r=.53 across grades). The weakest relationship was between holistic essay scores and posttest period composition grades (r=.13 across grades).

CHAPTER V

SUMMARY, CONCLUSIONS, AND IMPLICATIONS

Summary of Design

This study was designed to apply the holistic method of essay scoring at the school district level in order to directly measure writing ability and to evaluate a writing program. Holistic essay scoring has served as a reliable and efficient method for evaluating writing particularly in scoring essay components of national testing programs. The utility of the scoring procedure applied at the elementary school district level was analyzed by studying six questions.

1. Does the case study fit the theoretical model of holistic essay scoring?
2. Is consistency in scoring the essay papers achieved as determined by interrater reliability?
3. Is there a significant difference in test performance of seventh- and eighth-grade students as indicated by the obtained total test scores?
4. Is there significant growth in writing ability over a time frame of five months, from pretest to posttest?
5. Is the teacher in-service program effective as indicated by a significant difference in posttest scores obtained by the

treatment and no treatment groups?

6. Are the pretest and posttest holistic essay scores related to class composition grades and teacher ratings of student writing ability?

The data base for the study consisted of pre- and posttest essays written by 4,071 seventh- and eighth-grade students. Forty four teachers read and scored the essay examinations over a two-day period. Each paper received three holistic scores from three different readers. Two basic groups comprised the study sample; comparison (no treatment) group and experimental (treatment) group. Comparison group students were from classes whose teachers did not participate in the Writer's Clinic in-service program for teachers. These students did not respond to the pretest essay topic but did submit posttest essays. Experimental students' teachers did participate in the Writer's Clinic. The experimental group was comprised of three subgroups: pretest only; posttest only; and pretest/posttest matched pairs.

Additional data including student composition grades earned from December to May and teacher ratings of student writing ability as it was in December and as it was in May were supplied by fifteen of the forty four teachers involved in the holistic essay scoring session. These data were correlated with the pretest and posttest essay scores of 215 students.

Data were analyzed using analyses of variance, t-tests between means, alpha reliability coefficients, and Pearson correlation coefficients.

Findings

The case study applying the holistic method of essay scoring did

fit the theoretical model of the scoring technique. An approximation to the normal curve distribution was attained; 20.5 percent of the scores fell in the lower total score range (three to five), 15.9 percent fell in the higher end of the range (ten to twelve), and 63.7 percent of the scores fell in the middle range (six to nine). Obtained scores across each of the three readings per paper reflected reader reliability in scoring; reliability coefficients ranged from .77 to .84 for the essays within the various population subgroups. Eighth-grade students scored consistently higher than seventh-grade students on both the pretest and posttest. This was directly related to growth in writing ability with progression through the levels of education. Comparisons between pre- and posttest scores reflected higher scores at posttest time than at pretest time. Therefore, growth in writing, even over a five month instructional period, was substantiated. Determining the effectiveness of the Writer's Clinic program by comparing treatment and no treatment groups was not clearly established. This was due primarily to the fact that equivalency of the two groups was not certain (the comparison, no treatment group was not administered the pretest essay topic). Cautiously assuming equivalency would suggest the Writer's Clinic was effective; within the study subgroups, the experimental (treatment) group scored consistently higher than comparison (no treatment) group students on the posttest essay. Finally, teacher prepared data obtained for a sample of experimental group students, indicated a relationship between holistic essay scores and composition grades and teacher ratings of overall writing ability. The relationship was stronger between holistic essay scores and teacher ratings.

Conclusions

Within the limitations of the study, several conclusions are warranted. First, the holistic method used at the elementary school district level for evaluating student writing and a program designed to heighten that ability, seemed to be an efficient means of essay scoring. In a two-day period, 5,788 student essay examinations were each scored three times by forty four readers. While not as fast as machine scoreable multiple-choice test (indirect measurement) answer sheets, the readers essentially read 17,364 papers in approximately six hours; an average of about sixty four papers per hour per reader.

The ease of the scoring in terms of time, along with the relative reliability of the scoring, contributed to the efficiency and hence, utility of employing the holistic technique for the evaluation of writing. Holistic scoring as it was used in the study, provided a means of measurement that was able to discriminate among the good writers and the poorer writers. Hypothesized results, including a difference in performance of seventh-graders and eighth-graders and growth in writing over time were substantiated by the significant differences between grade levels and within pretest and posttest essay administrations.

Content and construct validity of the essay examination had been established prior to the essay scoring by clinicians of the Writer's Clinic program. The study was able to support some concurrent validity by finding relationships between the obtained essay scores and teacher ratings of writing ability. Procedures implemented for the scoring session confirmed the notion of ranking as did the teacher ratings; ranking in the sense of judging writing for what it is rather than for what it

should be. The scoring therefore, followed the theoretical model of holistic scoring; judging a paper against others rather than against a preconceived ideal.

Teacher acceptance of the holistic method, while not directly analyzed for this study, is of vital importance to the utility of the method. Utility of the method for the stated purpose of directly evaluating student composition was supported by the interest of the teacher group. The involved teachers, at first perhaps skeptical of the method, accepted it as a useful tool with additional applications for the classroom.

Educational Implications

The findings of this study indicate the holistic method of scoring provides one solution to the problem of evaluating writing directly, efficiently, and reliably. Holistic evaluation provides a starting point for evaluation. While readers may not be able to use the method for individual diagnoses, they are able to identify general problems in their students' writing. They see the writing as it is and can begin to view writing in its totality; to view writing as a whole rather than for individual aspects of writing skill.

Faced with impressions of weaknesses or strengths of student writing in general, the teacher can gear instructional objectives to improve students' abilities in weak areas as well as to build on the areas of strength. Looking at writing as it is, and working with students where difficulties lie, can bring the students closer to the preconceived ideal, closer to what writing truly should be.

Applications of the method for the classroom have few boundaries. The ease and quickness of the holistic method provides the teacher with a tool so that more writing assignments can be given in order to provide the student practice in writing and to provide the student the opportunity to express himself or herself by the written word.

Essay examinations will doubtfully ever replace multiple-choice tests of writing skill. Direct measures however, will continue to supplement indirect measures of writing ability. Likewise, holistic scoring can never replace, but can provide a supplement to, analytic scoring. And the total approach to analyzing writing can be of value not only for national standardized essay examinations but also for use within the individual classroom within and across subject area departments, within the school, and within the school district.

Implications for Research

Research in the area of directly measuring writing and applying the holistic method of essay scoring for that measurement is likely to continue. It will be recalled that in holistic scoring as implemented for this study, readers set standards for scoring and reached consensus concerning these standards through verbal discussion of training sample papers. In other words, written criteria or guidelines for scoring a paper in a certain way were not provided. It would therefore, be of interest to analyze what teachers of writing identify as attributes of those papers assigned high or low or middle scores. Such analyses would approach the evaluation of writing in an analytic sense. The comparison of analytic scoring and holistic scoring would be subject to study in this way.

It is a fundamental requirement in a modern society for its citizenry to possess a range of communication skills; writing being one of the most basic. Thus, continued research in assessing student writing skills is necessary to assist administrators and teachers in elementary and secondary education. The goal of such research is quite direct - to encourage thought and its expression.

REFERENCES

80,000 student essays graded in Atlantic City in December. College Board News, April 1978, p. 3.

Arnold, L. V. Writer's cramp and eyestrain - are they paying off? English Journal, January 1964, pp. 10-15.

Bowers, R. Progress Report on the Lilly Endowment Grant to the Indianapolis Public Schools for a Writer's Clinic. Unpublished report, January, 1978.

Bowers, R. Final Report on the Lilly Endowment Grant to the Indianapolis Public Schools for a Writer's Clinic. Unpublished report, September, 1978.

Braddock, R., Lloyd-Jones, R., & Schoer, L. Research in Written Composition. Urbana: National Council of Teachers of English, 1963.

Breland, H. M. A study of college English placement and the Test of Standard Written English (CEEB RDR-76-77, No. 4). Princeton: Educational Testing Service, 1977.

California State Department of Education. Evaluation improvement program: Program evaluator's guide. Princeton: Educational Testing Service, 1977.

Carroll, J. B. Note on the scoring of foreign language speaking and writing fluency tests (ETS RB 70-52). Princeton: Educational Testing Service, 1970.

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluative research: Strategies and methods. Pittsburg: American Institutes for Research, 1970.

Coffman, W. E. On the reliability of ratings of essay examinations in English. Research in the Teaching of English, 1971, 5, pp. 24-36.

Coffman, W. E. On the validity of essay tests of achievement. Journal of Educational Measurement, 1966, 3, pp. 151-156.

Cohen, A. M. Assessing college students' ability to write compositions. Research in the Teaching of English, 1973, 7, pp. 356-371.

Commission on English. Freedom and discipline in English. New York: College Entrance Examination Board, 1965.

Conlan, G. Essay and multiple-choice questions in tests of writing ability. Personal communication, March 24, 1978.

Conlan, G. How the essay in the CEEB English test is scored. Princeton: Educational Testing Service, 1976.

Conlan, G. Suggestions for writing essay questions. Personal communication, September 10, 1976.

Cooper, D. R. Measuring growth in writing. English Journal, 1975, 64, pp. 111-120.

Cooper, C. R. & Odell, L. Evaluating writing: Describing, measuring, judging. Buffalo: National Council of Teachers of English, 1977.

Cronbach, L. J. & Furby, L. How we should measure change - or should we? Psychological Bulletin, 1970, 74, pp. 68-80.

Diederich, P. B. Cooperative preparation and rating of essay tests. Princeton: Educational Testing Service.

Diederich, P. B. How to measure growth in writing ability. English Journal, 1966, 55, pp. 435-449.

Diederich, P. B. Measuring growth in English. Urbana: National Council of Teachers of English, 1974.

Diederich, P. B. Pitfalls in the measurement of gains in achievement. School Review, 1956, 64, pp. 59-63.

Diederich, P. B. & Link, F. R. Cooperative evaluation in English. In F. T. Wilhelms (Ed.) Evaluation as feedback and guide. Washington, D.C.: National Education Association, 1967.

Dressl, P., Kincaid, G., & Schmid, J. Journal of Educational Research, December 1952, pp. 285-293.

Educational Testing Service. Focus 5: The concern for writing. Princeton: Educational Testing Service, 1978.

Ebel, R. E. Essentials of educational measurement. Englewood Cliffs: Prentice-Hall, 1972.

Ebel, R. E. Estimation of the reliability of ratings. Psychometrika, 1951, 16, pp. 407-424.

Fagan, W. T., Cooper, C. R., & Jensen, J. M. Measures for research and evaluation in the English language arts. Urbana: National Council of Teachers of English, 1975.

Feinberg, S. G. Writing and the evaluating of writing at the post-secondary level. (Paper delivered to the English Department Composition Committee). Unpublished manuscript, Purdue University, 1978.

Follman, J. C. & Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 1, pp. 190-200.

Fowles, M. Basic Skills Assessment manual for the scoring of the writing sample. Princeton: Educational Testing Service, 1978.

French, J. W. Schools of thought in judging excellence of English themes. In Proceedings of the 1961 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1962.

Godshalk, F. I., Swineford, F., & Coffman, W.E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Hales, L. W. & Tokar, E. The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. Journal of Educational Measurement, 1975, 12, pp. 115-117.

Huddleston, E. Measurement of writing ability at the college entrance level: Objective vs. subjective techniques. Journal of Experimental Education, 1954, 22, pp. 165-213.

Hurtog, P. & Rhodes, E. C. The marks of examiners. New York: MacMillan, 1936.

Jerabek, R. & Diederich, D. Composition evaluation: The state of art. College Composition and Communication, 1975, 26, pp. 183-186.

Judine, Sister I. H. M., Ed. A guide for evaluating student composition. Urbana: National Council of Teachers of English, 1965.

Klein, S. P. & Hart, F. M. Chance and systematic factors affecting essay grades. Journal of Educational Measurement, 1968, 5, pp. 197-206.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in Measuring Change. Madison: University of Wisconsin Press, 1963.

Lord, F. M. The measurement of growth. Educational and Psychological Measurement, 1956, 16, pp. 421-437.

Marshall, J. C. & Powers, J. M. Writing neatness, composition errors, and essay grades. Journal of Educational Measurement, 1969, 6, pp. 97-101.

McColly, W. What does educational research say about the judging of writing ability? The Journal of Educational Research, 1970, 64, pp. 148-156.

Moslemi, M. H. The grading of creative writing essays. Research in the Teaching of English, 1975, 9, pp. 154-161.

Myers, A., McConville, C., & Coffman, W. E. Simplex structure in the grading of essay tests. Educational and Psychological Measurement, 1966, 26, pp. 41-54.

Nail, P., Fitch, R., Halverson, J., Grant, P., Winn, F. N. A scale for evaluation of high school student essays. Urbana: National Council of Teachers of English, 1960.

National Assessment of Educational Progress. Writing Report 3. Denver: National Assessment of Educational Progress, 1970.

National Assessment of Educational Progress. Writing Report 5. Denver: National Assessment of Educational Progress, 1971.

National Assessment of Educational Progress. Writing Report 8. Denver: National Assessment of Educational Progress, 1972.

National Assessment of Educational Progress. Writing Report 10. Denver: National Assessment of Educational Progress, 1972.

National Assessment of Educational Progress. Writing Report 11. Denver: National Assessment of Educational Progress, 1973.

National Assessment of Educational Progress. Writing Report 05-W-01. Denver: National Assessment of Educational Progress, 1975.

National Assessment of Educational Progress. Writing Report 05-W-02. Denver: National Assessment of Educational Progress, 1976.

Odell, L. The classroom teacher and researcher. English Journal, 1976, 65, pp. 106-111.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs: Prentice-Hall, 1978.

Sanders, S. E. & Littlefield, J. H. Perhaps test essays can reflect significant improvement in freshman composition: Report on a successful attempt. Research in the Teaching of English, 1975, 9, pp. 145-153.

Slotnick, H. B. Toward a theory of computer essay grading. Journal of Educational Measurement, 1972, 9, pp. 253-263.

Smith. R. Grading the Advanced Placement English Examination. Princeton: College Entrance Examination Board, 1976.

Smith, V. H. Measuring teacher judgement in the evaluation of written composition. Research in the Teaching of English, 1969, 3, pp. 181-195.

Stalnaker, J. M. & Stalnaker, R. C. Reliable reading of essay tests. The School Review, 1934, 8, pp. 599-603.

Tucker, L. R., Damarin, R., & Messick, S. J. A base-free measure of change. Psychometrika, 1966, 31, pp. 457-473.

Wagner, E. N. How to avoid grading compositions. English Journal, 1975, 3, pp. 76-79.

Walsh, M. F. Expressive writing in school: First year evaluation report. Berkeley: Educational Testing Service, 1978.

Whalen, T. E. A validation of the Smith Test for measuring teacher judgment of written composition. Education, 1972, 93, pp. 172-175.

Werts, C. E. & Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 73, pp. 17-22.

APPENDICES

APPENDIX A

THE ESSAY TOPIC AND

DIRECTIONS FOR ITS ADMINISTRATION

THE ESSAY TOPIC AND DIRECTIONS FOR ITS ADMINISTRATION

Write a theme no longer than one page in length answering the question: "Is television a help or hindrance to your getting a good education?" Remember to take a definite stand on the topic. Do not try to give two sides. Use an effective topic sentence in your opening paragraph. You may rephrase part of the question to form your topic sentence. You will have 30 minutes to write. At the end of that 30 minutes, I will collect your final draft. Your paper should be written for an adult reader.

In the controlled writing assignment:

1. There are no prewriting activities.
2. The assignment is given succinctly and a certain amount of time is allowed for writing.
3. Student compositions are written in ink on number five paper. Students should use the standard junior high school format including the school number next to the grade level.
4. Although students may revise and rewrite during the allotted time, teachers may not assist except to clarify the directions.
5. The assignments are not to be corrected or graded by the teacher. They are to be placed in manilla folders. Each folder should have the class number as well as the school number on it.

APPENDIX B

TRAINING SAMPLE PAPERS

TRAINING SAMPLE PAPERS

Score One

1. T.V. can hurt you because you could go blind if you sat to close to the T.V. and if you turn it up to loud it could hurt your ears and cause you to have alot of Headaches.

2. Television Is a Hindrance
To Getting an Education

Television is not Education to me because it dosn't tell anything about it and all it tells is love storys the storys are One Life to live, that dosn't say anything about school all the other shows like slave shows or other things like that it would be Educational because it is telling thing about the passed and the thing now is just talking about love if the show people just put on shows from the passed if thay can't the people can just try anyway.

3. Television is a hindrance

Television is a hindrance because. I never get any work done because. That all my mother let me do. Is whatch Televioion. When I was a little boy and now they are tring to stop me from whatching. It because. Televiosion When I was In chigao my uncle Thought that it was a hindrance to. Because his wife

(continued)

didn't even watch soap oprea I asked her she said I don't bevile in it. When I was out in chigao we just played chest road the byicle and whent to the baseball game and shot the bow and arrow

4. Is Television a Help or Hendrix.

I think television is both an help and a hendrix for example: I think it is good to be able to see countries and cultures different from ours, it is good to be able to know whats going on in countries besides ours.

Thanks to shows like Big Blue Marble, Call It Macaroni, and Make a Wish, that bring it to us.

But there are shows that are educational like Kojak, Barreta, and Switch. These are the shows that turn teen agers and adults etc.... into criminals. These shows teach them how to commit crimes ∴ Example: A kid is watching television and he see's the kid on television steal something so he says to himself that's easy, so the next day he goes out and gets caught, all because of television.

5. Television.

Television is okay. it will not mess you up. From getting an education. Well it helps me. Like these pictures when your best friends try to use you: You can find an way to stop them. Like on the picture

(continued)

called Millie. Well she had 4 friend and they tried to get her put out of school.

By framing her and By doing something that she wouldn't do. Like that day when Millie has an sub the girls wrote an topic using the sub's name adding awful words to it.

Well Millie solved her problem by doing the same thing they did to her. She did to them.

6. Television is a good education for me because there are a lot of show I like like good times if shows alot how to live without your mother and father. It hept you in all diferent kings of things.

Watch other shows help to. Television educational than some other things you do in life. It help you glow up better and teach you how to protect yourself and not let anyone bother you.

7. No, television is not a help because it is a bad influence to younger or older people who watch television because on tv like dective shows show a lot of violince and crime.

8. Television

Television helps me in wildlife shows like wild kindun, wourd of suvivile, gick crewstow shows with facts are my kind of shows

TRAINING SAMPLE PAPERS

Score Two

1. I think television is a hindrance to me.
Television is not help to me because there are too many cartoons or game shows on.
There need to be more educational television.

2. The violence on television

Television is a hindrance because it has alot of violence, but some people like violence on television which is alright with me. Some tv. shows have a whole lot of cussing on it. Which makes the younger kids cuss. It seems like kids know bad words before they know anything else. And they have all kinds of sex pictures on television which some mothers allow their kids to watch. And some do not allow their kids to watch. And the good tv. shows come on real late at night. Thats when I have to go to bed early so we can go to school. Pictures like Soap is a whole lot of violence to kids. Older people might like Soap. And the younger kids might like it too but their mother might not approve of them watching it. Because Parental discretion is advised. And that make kids forget about school because all they want to do is watch TV. They go to bed late then in the morning they falls to sleep in the room. Thats why television is a hindrance.

3. Television is a help for education. Because it knows alot more than teacher doese.
It teaches easer because you can't get distracted from a T.V.
And any time you can go get something to eat.
You can play and listen at the same time.

4. I think television is interfearing with me in getting a good education. cause almost every day after school I sit my homework on the dining room table and I go into my bedroom and watch television. but the next day when I go back to school I always have more work to do when I don't do my homework. So at the end of the 6 weeks when I get my report card my grades are always lower than my grades when I done my homework. And always when my grades are low my mother jumps on me and I get grounded from anything I want to do and I don't think that is why me watching television is interfearing with me getting a good education.

5. Television

I do think television is very educational. Because the word the people say are to hard for me to figure out. The words are very interesting though and the words are not little but big. Like cardiology

6. Television is helping my education

I like television because good movies comes on that I like. And the things they put on television like telling people not to smoke and the new when the man tells the weather when it is going to rain or snow.

I like the Christmas program they show and the police stories. I just like television because I watch It all the time.

7. Television is a help to my education

Television is a help to my education because I think all the shows I watch has some education in them.

I think all the shows put on the air are educational for everybody. Most of the shows teach you manners and things you should and shouldn't do. Most of the shows teach me things I don't already know.

8. TV

T.V. doesn't hurt me in school or either does it help me in school. I don't watch that much T.V. so there for it can't hurt me. And when I do watch T.V. it don't say anything that can help me in school. And mostly I only watch T.V. at night by that time I'll have my home work all done.

9. "Television"

"Television is a hindrance because of the children stay up and watch T.V. and cannot even get to school on

(continued)

time you know that is a shame. Some T.V. shows are violence like take Starsky & Hutch for instant some children stay up at 11:30 watching starsky and hutch. Then they come to school get into a fight and he remembers something off of a television they get to fight and will not stop until one shed's blood. It is just outrageous. really it don't make no since really it don't! Like Flip Wilson show he was always acting like someone he is not. This set a bad reputation for little children. they get out there in the street watching Flip Wilson acting Gerladine and can get hurt very bad it is just a shame. Mother's should limited television for small children.

Hindrance!

10. Televsion

I think televsion is health cause some shows are not vioatins like waltons they are drama and sadness. Police Story have to much vioatins anamal kingdom tells about nature and wildlife. Show shouldn't have killings in it or hating. baby Im back is love and hateful chips never firer there gun or chareles angle. Greese Adram is nature too. News tells about crim going down and what happing now adays. Happy Day is about Loving and caring and bradys bunch is to. Fintstone tells about in the cave man days. Jefferson hating and loving but they still is together to that why teleision is health

TRAINING SAMPLE PAPERS

Score Two/Three Split

1.

Is Television a help or a hindrance
in my getting a good education

I think Television is a help. Why? Because I like to look at Television when it is nothing to do. Like when it is cold out side you don't have nothing to do so you turn on the set.

At first it is a little Boring. And then it get a little better and then you just cant take your eyes off the set. I like most all the shows that come on. They have all kinds of shows. They have funny shows, silly shows they allso have sad shows on. But best

2.

Is Television A help or A hindrance
in my getting A good education?

Television is kind of a help, It teaches me way more than those trip owt teachers at school. They don't teach me nothin.

I learn way more on 30th street than at school. 30th street teaches a lot abowt the streets and teaches me abowt the politic's and math and whole bunch of junk, and the best thing abowt it you get good fights, to get yowr brain started up in the mroning, way better than a breakfast.

3.

Television - Good or bad?

Television at one time was great and was family

(continued)

entertainment. It used to cheer people up. In the 1930's with the big depression, it was a relief and enjoyable relaxation from the days Troubles. It was excitement, drama, nostalgia. There were horror movies & romance. But Today Television is all but good, it's trash. We need back The good ole days of Jackie Gleason, Red Skeleton, Laurel & Hardy, Charlie Chaplin, Abbott & Costello all The way To Dean Martin and Jerry Lewis - Even Though I wasn't born until "63", I've seen enough of today & yesterday To Know which I like. Television in my opionin is Terrible.

TRAINING SAMPLE PAPERS

Score Three

1. Yes, Television is a help to me, it makes me understand more about things that are going on in the world, it teaches me more about crime on the police shows and it tells me what they do about the crime.

The news shows tell me about what goes on around Indiana, and sometimes in other states or countries. I am glad that they thought up making the television, and I hope that it stays around.

2. Is Television A Help Or A Hindance?

Television is a help to our education.

Programs like Sesame Street, Electric Company and Janie encourage children to learn how to write, read, and speak. Children who watch these programs may be pre-educated by the time they start school.

News programs help people of all ages learn about different places of the world. Some of those shows would be: Toddy, 60 Minutes, Eye Witness News, and Good Morning America.

Everyone can benefit from television.

3. Television is a hindrance because when I'm doing my home work all I can hear is the television. My brothers always turn it up loud my little sister alway

(continued)

has her friends over and thier always screaming because they can't watch what they want. And the television station never shows nothing good except sports, and some movies. And when ever I'm doing my home work I always mess because I can't concentrate on my home work. So I just don't do it or I wait until the next morning and do it.

And then when I get home the next day it the same thing my little sister and her friends screaming and the television up too loud. So I just went into the bedroom and did my homework but I just barely get it done because I steal hear them screaming.

4. Television helps My Education

Television, to me is good for my education. Commercials helps me to understand words and what they mean. Movies also helps me to understand words, Television is really easy to learn off of, they put big words in little sentences. It helps me to learn what they're talking about. Television also teaches me how to use words. Television helps my education.

5. Televesion Education

Television is a help to my education. You can learn what is happening in other contries and let their problems help you trie to avoid making the same problem for your contry. Television also teaches parts of

(continued)

speach and some foran languages. It can help you learn the rools to games. When you need to relax before or after working you can watch tv. and you will be more relaxed and prepared for work. Some programs set examples so you would know more about what you would do in certain situations. It also teaches manners and how to act or what to do in certain kinds of places.

6. Television is help to me

It help me in so many ways. As one it tells me about the old times. It help with your school work, and when I need a friend I turn it on. It is like a guide some-times. I learn about crimes, and I learn about the city what go on. I learn about the weather, and actor's and singer's. I like television so much, I hope to be on television some day. I learn more about the race, Black & White, and etc. I learn about differnt animals, and other thing.

Television is a helpful thing to me.

7. Television Helps Crime

How many hours a week do you spend in front of the television?

If your like most kids your age you watch an average of 10 hours a week.

Almost every show on television today has alot of violens in it. This all gose in your head you tell

(continued)

your friends about it. Say how neat it was and say how much you would like to do it. Then someday you might try it and you could get caught.

All the crimes on television show step by step how to murder and rob someone. Telvision is a big hindranc to any kind of an education.

8. Television and Education

I think television is a help to education because programs on t.v. deal about school and education. While some programs help children know more about other places. It also helps people to lean about the cultures and the way people live. It also helps to ease the tension a child builds during the day at school. Sometimes it shows programs dealing with math and spelling. All the things is why I think T.V. is good for education.

TRAINING SAMPLE PAPERS

Score Three/Four Split

1. Education and Television

"Well your kid isn't learning because he watches too much T.V.," the teachers say to the parents. Well I say that's baloney. They say kids don't read because they're watching television. But if the kid doesn't like to read, he might be bored and become destructive. Then they say the student doesn't his homework because of T.V.. But if the kid doesn't want to learn, he won't.

In addition, T.V. might, through its documetaries, inspire a person to learn about a subject, which is what education is all about.

TRAINING SAMPLE PAPERS

Score Four

1.

Television and Education

Television I think is a hindrance to our education. There is too much violence on television, it disturbs children, they think they will grow up and be like "Barretta" or "Starsky and Hutch" or "Butch Cassidy" and the "Sundance Kid", even so I think this is not good for thier Education.

There should be education in television programs. There should be non-violent television shows. The kids shows even have violence, like the "Three Stooges" even some of the Flintstones shows have violence.

Sex on the telivision is outragous like "Noon till Three" and "The Spy who loved Me." these movies had sex almost all the way through it.

"Bonnie and Clyde" was a bad movie too for kids to watch, it had violence and sex. Robbory was what the movie was about and sex, and these two things set bad examples for the kids when they grow up.

2.

"T.V. is no Education"

Does television give you a good education, or is it bad on you?

I think television is bad on you. Because most

(continued)

of the shows you see have something to do with drugs, alcohol, fighting, or killing. When a little kid watches something like that it makes them want to do the same thing. They may say if they can do it, then I can to. On Starsky and Hutch there was a little boy on drugs and he said it was the only way to live. On Kojak there was this girl that ran away from home to be a hooker, and she didn't like school. There's a thousands of shows that come on like that or worser and it don't help our generation it just makes it worser. I think education would be a whole lot better if they took some of the shows like that off the air. How do you feel about television?

3. Television is a help to my getting a good education. Television is informative at most times and leads me to better understanding of the world around me.

Even when the T.V. shows are not basically educational, they still teach principles. Take, for instance, police shows they show a lot of violence but still have a point in them. They teach that crime is wrong.

I have not seen many shows yet that didn't have some moral. It is definent that T.V. will be an important part in the lives of people who want a vivid, pictorial view of life and education.

4. Television Isn't Bad

Television has many good points about it. T.V. has many educational shows. Shows like "Sesame Street" are shows that children can watch to learn to read, spell, and other skills.

Television has many shows on wildlife. When you watch one of these programs, you learn about animals far away or right close.

Some movies on T.V. are educational, too. Movies like "Midway" or "The Hindenberg" or even "How the West Was Won" are helpful in History classes.

One of the most educational shows that are on T.V. is the News. Local or national, it tells about what has been going on around the world. You can hear the president talk about inflation, or about a tragety.

I think television is one of the most helpful things in education today.

5. Television and Education

I think television helps me to get a better education. I learn that countries have different and interesting crafts and they also have different music, like ballet you could learn alot about ballet on television, like how to stand on your toes and turn around. Also you could learn about sports, like a country has a sport that you like and you couldn't see it knowhere else exept television. And you could also

(continued)

learn about the different problems others countries are having. So I think television helps alot espically if you want to learn and enjoy televisions help.

6. Television

I think television is a hinderance to my getting a good education because of all of the violence on television. You watch a program and then you go to do your work, but all that your thinking about is that program so you don't think about your work.

Television is often a hinderance when you watch a show one night then you go to school. You're thinking about that show then everyone else starts talking about it. You and every one who saw the show would want to act like the characters in the show and forget all about your work. Pretty soon you'll be sitting down at the principal's office ready for a paddling.

So do you see why I think television is a hinderance to my getting a good education?

APPENDIX C

MASTER SCORING CODE

HOLISTIC ESSAY SCORING SESSION

May 10-11, 1978

Indianapolis, Indiana

MASTER CODE

4.	B	E	F	K	P	R	
3.	I	J	L	S	T	Z	
2.	M	N	V	W	X	Y	U
1.	A	C	G	H	O	Q	D

20.	BIMA	30.	KTYD	40.	EJVG	50.	RTXO	60.	BJUC
21.	EJNC	31.	RTWG	41.	FLMH	51.	EIMA	61.	EIYO
22.	FLVG	32.	PSVC	42.	KSXO	52.	FJNC	62.	FJMQ
23.	KSWH	33.	KLWA	43.	PTYQ	53.	KLYG	63.	RING
24.	PTXO	34.	BLMG	44.	RZUD	54.	RLWH	64.	KJWD
25.	RZYQ	35.	ESND	45.	BJMC	55.	RIMA	65.	PLUA
26.	BIUD	36.	FTVO	46.	FSVH	56.	PZYQ	66.	BSVA
27.	BJVH	37.	KZXQ	47.	KTWO	57.	KTXO	67.	EJMA
28.	ELWO	38.	PZUC	48.	PZXQ	58.	FSWH	68.	PIWG
29.	FSXQ	39.	BINC	49.	EZNA	59.	ELVD	69.	FSNO

APPENDIX D

CROSSTABULATIONS OF FIRST AND SECOND READING SCORES CONTROLLING FOR THIRD READING VALUES

Table 59

Crosstabulation of Comparison Group Posttest Essay Scores for The First and Second Reading, Controlling For Third Reading Value of 1

First Reading Scores	Second Reading Scores 1	2	3	Row Total
1	27*	14	0	41 60.3
2	7	16	2	25 36.8
3	0	2	0	2 2.9
Column Total	34 50.0	32 47.1	2 2.9	68 100.0

Correlation = .42

*Perfect agreement among three readings.

Table 60

Crosstabulation of Comparison Group Posttest Essay Scores for The First and Second Reading Controlling For Third Reading Value of 2

First Reading Scores	Second Reading Scores 1	2	3	4	Row Total
1	21	35	7	0	63 18.1
2	24	139*	41	0	204 58.6
3	4	30	36	4	74 21.3
4	0	0	7	0	7 2.0
Column Total	49 14.1	204 58.6	91 26.1	4 1.1	348 100.0

Correlation = .41

*Perfect agreement among the three readings.

Table 61

Crosstabulation of Comparison Group Posttest Essay Scores for The First and Second Reading Controlling For Third Reading Value of 3

First Reading Scores	Second Reading Scores 1	2	3	4	Row Total
1	0	5	0	0	5 2.6
2	3	40	37	2	82 42.9
3	0	29	40*	8	77 40.3
4	0	7	15	5	27 14.1
Column Total	3 1.6	81 42.4	92 48.2	15 7.9	191 100.0

Correlation = .30

*Perfect agreement among the three readings.

Table 62

Crosstabulation of Comparison Group Posttest Essay Scores for The First and Second Reading Controlling For Third Reading Value of 4

First Reading Scores	Second Reading Scores 2	3	4	Row Total
2	0	11	0	11 26.8
3	3	7	5	15 36.6
4	0	6	9*	15 36.6
Column Total	3 7.3	24 58.5	14 34.1	41 100.0

Correlation = .42

*Perfect agreement among the three readings.

Table 63

Crosstabulation of Experimental Pretest/Posttest Matched Pairs Subgroup Pretest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 1

		Second Reading Scores			
Count		1	2	3	Row Total
First Reading Scores	1	64*	36	0	100 48.3
	2	30	59	7	96 46.4
	3	0	11	0	11 5.3
Column Total		94 45.4	106 51.2	7 3.4	207 100.0

Correlation = .39

*Perfect agreement among the three readings.

Table 64

Crosstabulation of Experimental Pretest/Posttest Matched Pairs Subgroup Pretest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 2

		Second Reading Scores				
Count		1	2	3	4	Row Total
First Reading Scores	1	41	101	19	0	161 19.2
	2	62	294*	103	0	459 54.6
	3	8	113	74	8	203 24.2
	4	0	0	17	0	17 2.0
Column Total		111 13.2	508 60.5	213 25.4	8 1.0	840 100.0

Correlation = .34

*Perfect agreement among the three readings.

Table 65

Crosstabulation of Experimental Pretest/Posttest Matched Pairs Subgroup Pretest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 3

First Reading Scores	Second Reading Scores 1	2	3	4	Row Total
1	0	12	0	0	12 2.2
2	7	87	78	9	181 33.1
3	0	76	132*	36	244 44.6
4	0	19	52	39	110 20.1
Column Total	7 1.3	194 35.5	262 47.9	84 15.4	547 100.0

Correlation = .38

*Perfect agreement among the three readings

Table 66

Crosstabulation of Experimental Pretest/Posttest Matched Pairs Subgroup Pretest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 4

First Reading Scores	Second Reading Scores 2	3	4	Row Total
2	0	7	0	7 5.7
3	6	23	13	42 34.1
4	0	20	54*	74 60.2
Column Total	6 4.9	50 40.7	67 54.5	123 100.0

Correlation = .48

*Perfect agreement among the three readings.

Table 67

Crosstabulation of Experimental Pretest/Posttest Matched Pairs Subgroup Posttest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 1

First Reading Scores	Count	Second Reading Scores 1	2	3	Row Total
	1	51*	21	0	72 53.7
	2	19	31	2	52 38.8
	3	0	10	0	10 7.5
	Column Total	70 52.2	62 46.3	2 1.5	134 100.0

Correlation = .44

*Perfect agreement among the three readings.

Table 68

Crosstabulation of Experimental Pretest/Posttest Matched Pairs Subgroup Posttest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 2

First Reading Scores	Count	Second Reading Scores 1	2	3	4	Row Total
	1	29	73	17	0	119 15.9
	2	45	238*	105	0	388 51.9
	3	8	104	92	12	216 28.9
	4	0	0	24	0	24 3.2
	Column Total	82 11.0	415 55.6	238 31.9	12 1.6	747 100.0

Correlation = .37

*Perfect agreement among the three readings.

Table 69

Crosstabulation of Experimental Pretest/Posttest Matched Pairs Subgroup Posttest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 3

First Reading Scores	Second Reading Scores 1	2	3	4	Row Total
1	0	11	0	0	11 1.7
2	8	89	76	10	183 28.5
3	0	63	196*	54	313 48.7
4	0	20	71	45	136 21.2
Column Total	8 1.2	183 28.5	343 53.3	109 17.0	643 100.0

Correlation = .40

*Perfect agreement among the three readings.

Table 70

Crosstabulation of Experimental Pretest/Posttest Matched Pairs Subgroup Posttest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 4

First Reading Scores	Second Reading Scores 2	3	4	Row Total
2	0	11	0	11 5.7
3	7	31	24	62 32.1
4	0	34	86*	120 62.2
Column Total	7 3.6	76 39.4	110 57.0	193 100.0

Correlation = .42

*Perfect agreement among the three readings.

Table 71

Crosstabulation of Experimental Pretest Only Subgroup Pretest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 1

		Second Reading Scores			
	Count	1	2	3	Row Total
First Reading Scores	1	41*	21	0	62 55.9
	2	21	17	6	44 39.6
	3	0	5	0	5 4.5
	Column Total	62 55.9	43 38.7	6 5.4	

Correlation = .32

*Perfect agreement among the three readings.

Table 72

Crosstabulation of Experimental Pretest Only Subgroup Pretest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 2

		Second Reading Scores				
	Count	1	2	3	4	Row Total
First Reading Scores	1	20	57	7	0	84 21.4
	2	26	143*	45	0	214 54.6
	3	3	39	33	6	81 20.7
	4	0	0	13	0	13 3.3
	Column Total	49 12.5	239 61.0	98 25.0	6 1.5	

Correlation = .42

*Perfect agreement among the three readings.

Table 73

Crosstabulation of Experimental Pretest Only Subgroup Pretest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 3

First Reading Scores	Count	Second Reading Scores 1	2	3	4	Row Total
	1	0	6	0	0	6 2.4
	2	5	37	37	2	81 32.8
	3	0	30	65*	12	107 43.3
	4	0	9	28	16	53 21.5
	Column Total	5 2.0	82 33.2	130 52.6	30 12.1	

Correlation = .41

*Perfect agreement among the three readings.

Table 74

Crosstabulation of Experimental Pretest Only Subgroup Pretest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 4

First Reading Scores	Count	Second Reading Scores 2	3	4	Row Total
	2	0	4	0	4 7.4
	3	3	12*	11	26 48.1
	4	0	9	15	24 44.4
	Column Total	3 5.6	25 46.3	26 48.1	

Correlation = .33

*Perfect agreement among the three readings.

Table 75

Crosstabulation of Experimental Posttest Only Subgroup Posttest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 1

		Second Reading Scores			
	Count	1	2	3	Row Total
First Reading Scores	1	25*	22	0	47 49.5
	2	13	27	4	44 46.3
	3	0	4	0	4 4.2
	Column Total	38 40.0	53 55.8	4 4.2	95 100.0

Correlation = .31

*Perfect agreement among the three readings.

Table 76

Crosstabulation of Experimental Posttest Only Subgroup Posttest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 2

		Second Reading Scores				
	Count	1	2	3	4	Row Total
First Reading Scores	1	19	39	7	0	65 16.0
	2	28	130*	46	0	204 50.2
	3	3	57	52	5	117 28.8
	4	0	0	20	0	20 4.9
	Column Total	50 12.3	226 56.7	125 30.8	5 1.2	406 100.0

Correlation = .44

*Perfect agreement among the three readings.

Table 77

Crosstabulation of Experimental Posttest Only Subgroup Posttest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 3

		Second Reading Scores				
	Count	1	2	3	4	Row Total
First Reading Scores	1	0	6	0	0	6 2.1
	2	2	37	42	4	85 29.2
	3	0	38	75*	23	136 46.7
	4	0	11	23	30	64 22.0
	Column Total	2 0.7	92 31.6	140 48.1	57 19.6	219 100.0

Correlation = .40

*Perfect agreement among the three readings.

Table 78

Crosstabulation of Experimental Posttest Only Subgroup Posttest Essay Scores for the First and Second Reading Controlling for Third Reading Value of 4

		Second Reading Scores			
	Count	2	3	4	Row Total
First Reading Scores	2	0	6	0	6 5.5
	3	2	19	15	36 32.7
	4	0	10	58*	68 61.8
	Column Total	2 1.8	35 31.8	73 66.4	

Correlation = .53

*Perfect agreement among the three readings.

APPROVAL SHEET

The thesis submitted by Judith A. Powills has been read and approved by the following committee:

Dr. Jack A. Kavanagh
Associate Professor and Chairman, Education Foundations

Dr. Ronald R. Morgan
Assistant Professor, Education Foundations

The final copies have been examined by the director of the thesis and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the thesis is now given final approval by the Committee with reference to content and form.

The thesis is therefore accepted in partial fulfillment of the requirements for the degree of Master of Arts.

April 20, 1979
Date

Jack A. Kavanagh
Director's Signature